AD-A033 920 ARIZONA STATE UNIV TEMPE COLL OF EDUCATION F/G 5/9
AN AUTOMATED MEASUREMENT TECHNIQUE FOR EVALUATING PILOT SKILL.(U)
FEB 76 B D SHIPLEY AF-AFOSR-2900-76
UNCLASSIFIED TN-60229 AFOSR-TR-76-1253 NL

1 OF 2
ADA033920

ADA033920

AFOSR - TR - 76 - 1253

STUDIES IN SYSTEMATIC INSTRUCTION AND TRAINING

12

# AN AUTOMATED MEASUREMENT TECHNIQUE FOR EVALUATING PILOT SKILL


Brian D. Shipley





USAF Office of Scientific Research

Grant No. 76-2900


ARIZONA STATE UNIVERSITY
EDUCATIONAL TECHNOLOGY

coll. of Education
409989



Technical Note #60229

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

REPORT DOCUMENTATION PAGE

1. REPORT NUMBER

AFOSR - TR - 76 -

4. TITLE (and Subtitle)

(6) An Automated Measurement Technique for Evaluating Pilot Skill.

5. TYPE OF REPORT & PERIOD COVERED

(9) Interim rept.

6. PERFORMING ORG. REPORT NUMBER

(14) TN-60229

7. AUTHOR(s)

(10) Brian D. Shipley

8. CONTRACT OR GRANT NUMBER(s)

(15) AF-AFOSR-2900-76

9. PERFORMING ORGANIZATION NAME AND ADDRESS

College of Education
Arizona State University
Tempe, AZ 85281

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

(16) 61102F/9778/681313

11. CONTROLLING OFFICE NAME AND ADDRESS

A. F. Office of Scientific Research (NL)
Bolling AFB, DC 20332

12. REPORT DATE

(11) February, 1976

13. NUMBER OF PAGES

120 (12) 132p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

(18) AFOSR

(19) TR-76-1253

15. SECURITY CLASS. (of this report)

Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Performance skill
Pilot training
Evaluation of skill
Evaluation model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Specific indicators of performance skill in pilot training are lacking. This dissertation represents an effort to begin filling that void. An algorithmic, performance state evaluation model was developed for an instrument flight maneuver with performance times and deviations from a standard flight path as indicators of skill. The algorithm's initial procedures and these indicators were used in three empirical investigations. The first investigation showed that performance times can be used to enable an observer to discriminate between per-
(Continued on back)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED 409 989

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

formances or performance states in performances by two experienced pilots. In the second investigation, means of total performance time were found to discriminate between differences in treatments used in a training experiment with student pilots as subjects (data from Brecke, 1975); an a priori predictions of differences in effects of these treatments on variability of group performances at a specified location were significant.

In the final investigation, support was found for the hypothesis that a small set of specific indicators could be used to replace a summary indicator of variability in performances. Results of stepwise regression analyses indicated that 7 of 12 specific indicators could be used to account for 34% to 82% of the variance in the summary indicator over 6 performance trials and that there were nearly identical curvilinear trends in means ($r$ = .98) due to improvement over trialsfor the summary indicator and deviations from a standard. It was concluded that with the present maneuver, the model allowed for superior evaluations with fewer data points. The need to test detailed analytic procedures in the model and to extend the methods used in the development of the model to other pilot training maneuvers was discussed.

AN AUTOMATED MEASUREMENT TECHNIQUE
FOR EVALUATING PILOT SKILL

Rule Learning and Systematic Instruction in
Undergraduate Pilot Training

Vernon S. Gerlach, Principal Investigator

# AN AUTOMATED MEASUREMENT TECHNIQUE FOR EVALUATING PILOT SKILL

Brian D. Shipley

Technical Note #60229

Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR 76-2900. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

College of Education
Arizona State University
Tempe, Arizona

February, 1976

## ABSTRACT

This dissertation seeks to discover

Specific indicators of performance skill in pilot training are lacking. This dissertation represents an effort to begin filling that void. An algorithmic, performance state evaluation model was developed for an instrument flight maneuver with performance times and deviations from a standard flight path as indicators of skill. The algorithm's initial procedures and these indicators were used in (3) ~~three~~ empirical investigations. The first investigation showed that performance times can be used to enable an observer to discriminate between performances or performance states in performances by two experienced pilots. In the ~~second~~ 2nd investigation, means of total performance time were found to discriminate between differences in treatments used in a training experiment with student pilots as subjects (data from Brecke, 1975); and *a priori* predictions of differences in effects of these treatments on variability of group performances at a specified location were significant.

In the final investigation, support was found for the hypothesis that a small set of specific indicators could be used to replace a summary indicator of variability in performances. Results of stepwise regression analyses indicated that 7 of 12 specific indicators could be used to account for 34% to 82% of the variance in the summary indicator over 6 performance trials, and that there were nearly identical curvilinear trends in means ($r$ = .98) due to improvement over trials

for the summary indicator and deviations from a standard. It was concluded that with the present maneuver, the model allowed for superior evaluations with fewer data points. The need to test detailed analytic procedures in the model and to extend the methods used in the development of the model to other pilot training maneuvers was discussed.

TABLE OF CONTENTS

LIST OF TABLES

| Table | | Page |
|---|---|---|

## LIST OF FIGURES

blank

# CHAPTER I

## THE PROBLEM OF PERFORMANCE EVALUATION IN PILOT TRAINING

This chapter is an introduction to the problem of measurement and evaluation of skill in pilot training. In pilot training, evaluations are used to diagnose student pilot learning difficulties, to manage the training program, and to conduct training research and development. Because pilots operate a complex system in an unstable, frequently dangerous environment, problems result from differences among evaluation methods needed and used in the operational, the management, and the research settings. To be fully effective, any formal measurement and evaluation methods must ultimately be usable in each of these settings. This study was designed to investigate problems of measuring and evaluating skill in pilot performances from the view of training research and development.

The thesis of this study was that specific indicators of skill in pilot performances could be used with a performance state evaluation model to resolve a measurement dilemma: excessive detail versus uninformative generality. To establish a general framework for this thesis, the problems of evaluation which confront the instructor pilot are described first. This description is followed by a delineation of problems with the evaluation methods generally used in pilot training management. Next, the needs for evaluation methods in pilot training research and development are considered. Finally, the major

points are summarized and the specific purposes of this study are stated.

## Instructor Pilot Evaluation Problems

Pilots operate a complex system in an unstable, frequently dangerous environment. Consider how these operational factors make demands on an instructor pilot (IP) as he evaluates a student pilot's performance in an aircraft. The IP must monitor the student pilot's (SP) behavior, the aircraft, and the surrounding airspace. He must identify inappropriate SP behaviors, unsafe performance conditions, and dangers in the airspace. When he detects an error or threatening condition, he must decide what action will best meet the objectives of mission safety and the training needs of the SP. He must carry out the selected action. Observations in all these areas of performance must be remembered or recorded and then used to arrive at a meaningful score for the performance.

Clearly, the IP must process large quantities of information to fulfill the requirements of his assignment. As training tasks become more complex or dangerous, the demands on the IP tend to increase and IPs will tend to be less and less able to adequately process all the essential information. If the information processing load on the IP becomes excessive, the integrity of safety procedures, training effectiveness, and evaluation methods will be compromised. These requirements and conditions have resulted in general use of rating methods to evaluate skill in performances during pilot training.

## Evaluation Methods in Pilot Training

After more than 30 years, rating scales are still the basis of evaluation methods in pilot training. Rating scales are used because, as yet, there are no cost effective evaluation methods to meet the needs of the operational user, the IP (Koonce, 1974). Some alternatives to rating scale methods have been investigated. These alternatives are paper and pencil observation schedules and, more recently, automated data collection and computer aided evaluation systems.

Rating scales are still used because they meet the needs of training management without intruding seriously into the IPs' operational capabilities.

As currently used, rating scale methods are not adequate to meet the needs of training research and development. Rating scale methods are inadequate for these needs because observations using them lack sufficient discrimination, i.e., the observations tend to accumulate on one or two points in the rating scale. Without better discrimination, rating methods are not adequate as criteria to validate alternative methods of measurement and evaluation (Knoop & Welde, 1973). To make IP observations more discriminating in training research and development, extensive observer training and quality control programs must be developed and carried out (Horner, Radinsky, & Fitzpatrick, 1970; Koonce, 1974).

In their present form, IP ratings cannot be used in training research and development. Consider the problems of standardization of IP ratings under the present system. Inflight pilot training and

proficiency evaluation is largely on a one-to-one basis; this means that pilot instructors train IPs and check pilots who, in turn, train and evaluate SPs. Since training personnel cannot simultaneously observe and evaluate an SP's performance, they have little common basis for developing uniform criteria. Rather, their training and evaluation skills evolve primarily from their personal experiences in the highly complex aircraft environment. Measures and criteria developed in training research might be used to improve the effectiveness of IP ratings.

Christensen and Mills (1967) contended that methods for the evaluation of complex performances suffer from a criterion problem. Quoting Thorndike (1947, p. 29), these authors noted that the criterion problem is one "of obtaining satisfactory criterion measures against which to validate tests and evaluate variations of training methods" (Christensen & Mills, 1967, p. 335). Observations from recent pilot training evaluation studies confirm this observation (Horner, et al., 1970; Knoop & Welde, 1973). Christensen and Mills also argued that quantification in the absence of a criterion is wasted effort: "to generate numbers does not automatically assure either understanding or validity" (p. 335). As a conclusion these authors observed that "we seem to be no nearer than we were 20 years ago to the development of independent, uncontaminated criteria of human performance under operational conditions" (p. 339).

Seven years later, Koonce (1974) reviewed problems of measuring and evaluating pilot performances. A criterion problem still existed. He found that measures derived from a sound theoretical position did

not exist. Rather, he found that trial and error methods were used with data from every possible measurable source, or alternatively, that methods of systematic inspection and expert judgment were used to select measures for particular effects. Training research and development efforts cannot be effective without valid response measures and consistent evaluative criteria. Perhaps the problems of measurement and evaluation validity ought to be viewed in a different way.

## Evaluation Methods in Pilot Training Research

Measures and criteria used in pilot training research and development might be used as a basis to improve IP evaluation methods. In training research and development, training objectives, instruction, and evaluation are three interrelated components. Generally, piloting skills and competencies, to be developed in training, are identified in the objectives. Instruction covering these competencies is prepared and SPs are asked to use this instruction to perform assigned tasks under specified test conditions.

Iterative procedures are used in training research and development: SP performances are observed and "if the student's responses do not correspond with the specified outcomes, the materials are revised and the process is repeated" (Merrill, 1971, p. 2). In this iterative process, it is generally assumed that the necessary response measures and criteria exist and that relationships between measures and criteria are well known. Evidently, such assumptions must be questioned in the area of pilot training (Brecke & Gerlach, 1972). It follows

that effective pilot training research efforts must include an investigation of measures of skill in the area of interest (Koonce, 1974; Shipley, Gerlach, & Brecke, 1974). The results of these and related investigations should be studied to improve our understanding of measures of skill and methods of evaluation in pilot training.

## Summary

Piloting performances are complex because they are composed of many events. Some of these events will be more crucial to the attainment of mission objectives than others. Critical events probably occur at different times or places during a performance but at similar times and places for repeated performances of the same task. For the purposes of measurement and evaluation of skill in pilot training, it would be helpful to know the characteristics of these critical events. For a given task, it should be possible to determine empirically such factors as the relative importance and the time or placement of such events in the operational sequence. Given such empirical data, an evaluator would have a basis to develop objective performance measurement and evaluation standards.

A training research approach to measurement and evaluation problems in pilot training should include these objectives:

1. To identify potentially critical points or events in descriptions of pilot behaviors that make up the operational sequence of the performance tasks.

2. To develop observation schedules and scoring procedures to account for the effects of these events on performance skill.

3. To determine empirically relative frequencies of these critical events throughout performances of the assigned task from objective data.

4. To train IPs, check pilots, and other pilot training personnel to employ the schedules and procedures with SP performances, first in a simulator, then in the aircraft.

5. To develop reliability assessment procedures for use with measurement and evaluation practices in the aircraft based on the outcomes of the four preceding objectives.

In the present study, the first three objectives were investigated from the view of a pilot training research and development project. A secondary purpose was to investigate measures and methods that might be usable in the operational environment. The primary purpose was to identify or develop indicators of performance skill that would resolve a recurring dilemma in pilot training measurement and evaluation: excessive detail versus uninformative generality (Youtz & Ericksen, 1947). The hypothesis was that a set of specific indicators could be used to replace a set of more general, summary indicators in training research and development. In the context of the present study, specific indicators were measures of performance events at particular points in a performance. Alternatively, summary indicators were measures obtained as a function of sums computed from each observation in an entire set of time series data from the same performance.

# CHAPTER II

## THEORETICAL ANALYSIS AND REVIEW OF LITERATURE

In this chapter, a theoretical analysis is used to define a dependent variable, precision of control, and to relate it to characteristics of skill in piloting performances. This analysis is followed by a review of the pilot training literature on selected measurement and evaluation methods. The objective of the review was to evaluate measures and methods that might be used as indicators of precision of control. The first two sections of the chapter contain the results of the theoretical analysis. In the third section, response measures and methods are reviewed. In the fourth section, a simple performance state evaluation model is described. The last section is a summary of the chapter.

In this chapter, the concept of a "state" from control theory is used to refer to segments of a flight path. It is assumed that training objectives designate the characteristics of the flight path to be used in student performances. Referring to Figure 1, these training objectives prescribe the system inputs to the pilot and define the characteristics of the flight path expected as system performance outputs. As used in the present study, the evaluation model is essentially a process of comparison between specified and measured performances.

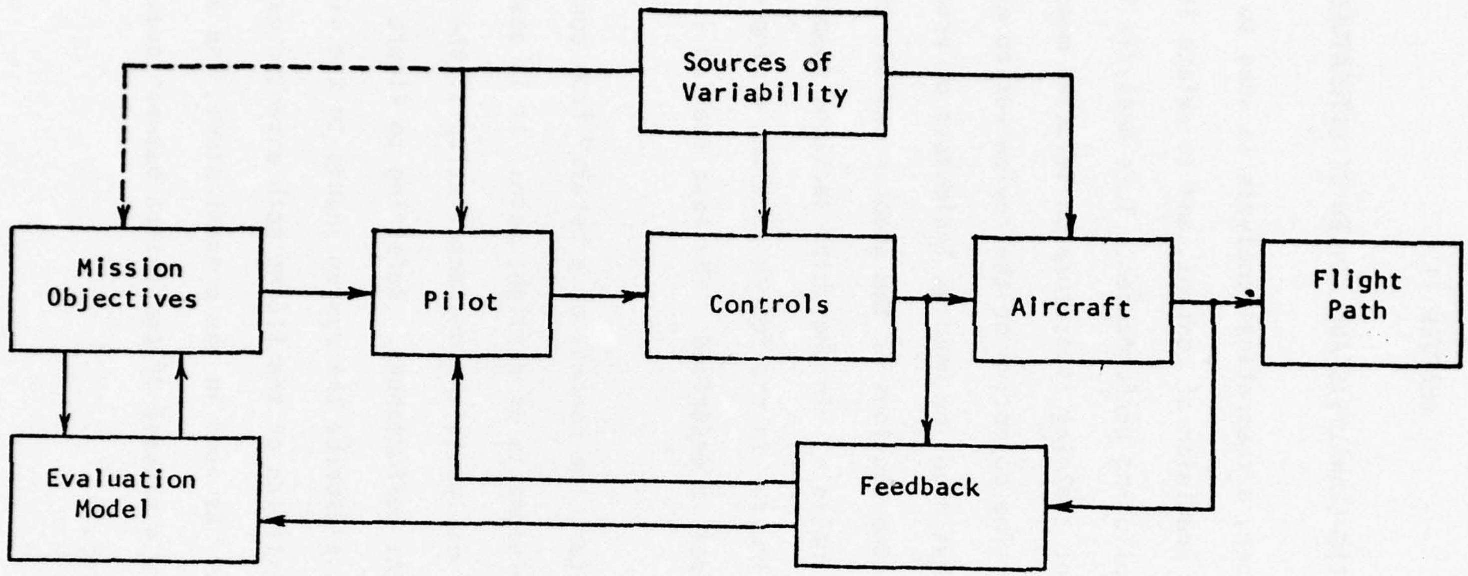


Figure 1.--Model of Pilot-Aircraft Performance System

As used here, "state" is a measurement or output determined concept. This simple application of "state" differs somewhat from the sense of more formal definitions in which "state" is used to represent the internal workings of some dynamic system, i.e., the pilot (Figure 1). In previous research, Connelly, Schuler, and Knoop (1969) began with a similar definition of a state. However, their purpose was to mathematically simulate the behavior of a human evaluator. To do this modeling operationally, they used adaptive mathematical methods on a computer without any prior assumptions about relevant performance measures or criteria, i.e., inputs or system performance variables.

## Describing Complex Performances

Uncertainty is a part of every pilot's, every instructor's, and every evaluator's mission in pilot training. An aircraft's flight path is not visible and a pilot's control behaviors leave no permanent trace. Air is unstable and weather factors cause unpredictable variations in the performance. Changes in a flight path will also occur because the airspace must be shared with other aircraft and because obstacles must be avoided over mountainous terrain or at low altitudes. Pilots, instructors, and evaluators also possess the attribute, "unpredictability." To describe a flight path under any of these conditions, an evaluator has been faced with a choice between unmanageable detail on the one hand or uninformative generalities on the other (Youtz & Ericksen, 1947).

In the pilot-aircraft system one of the pilot's functions is to control the aircraft (Figure 1). The results of this control are

reflected through feedback as values on the aircraft instruments and as perceptual cues. An example of a visual cue is the orientation and location of the horizon relative to the aircraft altitude and attitude. A pilot uses these values and cues as feedback to make decisions about any changes in the controls that are needed to accomplish his mission. Mission objectives, control changes, instrument values, and cues can be systematically related to each other through the principles of control theory. These principles are a useful tool in specifying values and cues in a standard training maneuver, in determining a list of control movements essential to complete a prescribed flight path, and in the development of evaluation methods (Brecke & Gerlach, 1972; Gerlach, Brecke, Reiser, & Shipley, 1972; Shipley, Gerlach, & Brecke, 1974).

To approximate a standard flight path as he executes a maneuver, a pilot uses information from instruments, references outside the aircraft, and the mission objectives. The precision of control exhibited by the pilot in maintaining a flight path will vary with such factors as his alertness, his level of piloting skill, the varying requirements of the mission, and differences in flight activities. For example, in the transition from climb to descent in the Vertical S-A (Figure 2), variability in maximum altitude was found to discriminate between groups of student pilots given different sets of preflight instructions (Brecke, Gerlach, & Shipley, 1974). Precision of control is an important dependent variable to consider in evaluating pilot performances.

The correctness of a pilot's control of the aircraft is evaluated by how well it approximates a "standard." The standard should be

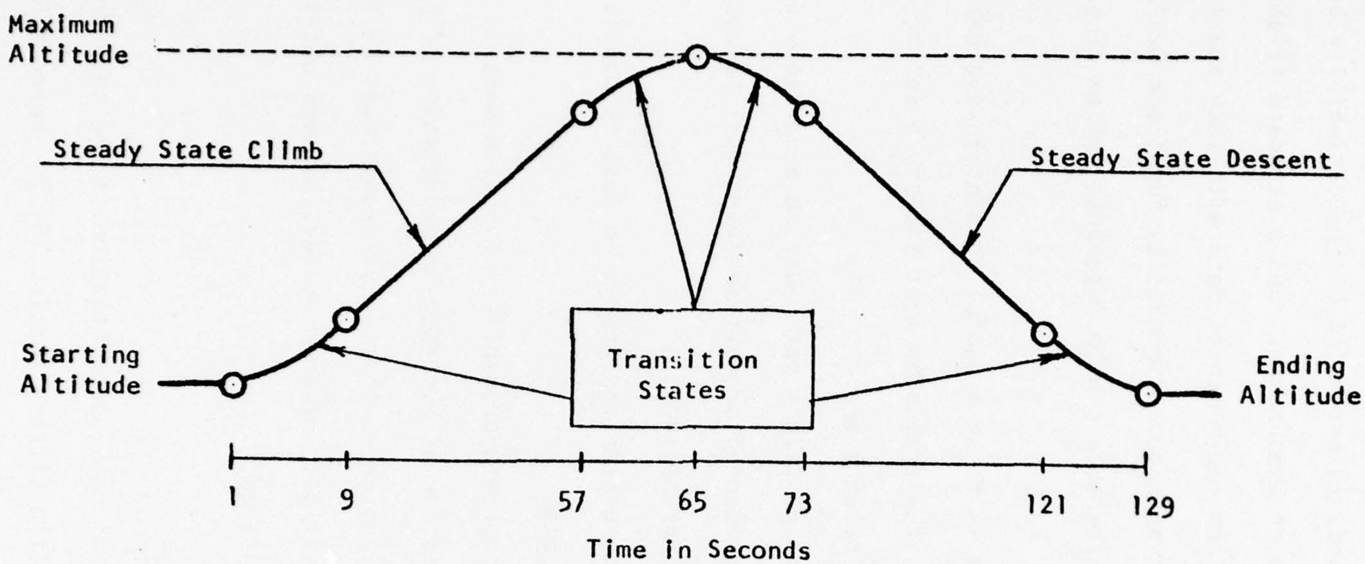


Figure 2.--Standard Vertical S-A Altitude Curve Showing Performance States and Performance Times

specified in the performance objectives for a mission. Precision, a concept which expresses the degree of correspondence between an observation and a standard, is estimated by the variability between experimental observations or measurements and a standard flight path. Precision is estimated from experimental data with such statistics as the variance and the covariance. Probability functions applied to these statistical measures enable one to interpret degree of precision in flight path data.

If precision is to be a useful concept in evaluating student pilot performances, the evaluator must satisfy a set of basic requirements. These requirements are:

1. The standard flight path must be accurately defined.
2. The experimental or test maneuver must be accurately observed or measured.
3. Rules for estimating precision must be consistently applied to the data.
4. The appropriate probability function must be applied to the estimates for an accurate interpretation of precision.

Methods for satisfying these four requirements must be identified or developed and the validity of any methods, however obtained, must be empirically demonstrated.

## Standard Flight Path

In the general case, specification of the standard flight path for test purposes is arbitrary (Etkin, 1972). Horner, Radinsky, and Fitzpatrick (1970) reviewed descriptions of the standard training

maneuvers used in the Air Force Undergraduate Pilot Training (UPT) syllabus. They concluded that ideal or standard flight path values were adequately specified. That is, they were able to define rating scales for instructor pilots to evaluate video-taped performances.

Knoop and Welde (1973) arrived at a different conclusion. In an investigation of automated data collection and evaluation (ADCS) methods, they found that several descriptions of UPT maneuvers were inadequate for developing computerized evaluation methods. In some cases the descriptions contained insufficient information for developing the necessary equations. In other cases, performances of experienced instructor pilots were found to differ in form from the standard flight path. Finally, they found that instructor pilots were unable to perform some of the maneuvers as specified because the standard flight paths were beyond the performance limits of the pilot-aircraft system. Knoop and Welde (1973) concluded that empirical methods must be used to support a control theory analysis of descriptions given in manuals and textbooks or by instructor pilots.

## Application of Control Theory

### Performance States

Pilot training researchers or performance evaluators can use control theory as an aid in understanding the concept of a standard flight path. One can describe movements of an aircraft through space and time (i.e., the flight path) with values from several variables. Examples of these variables include among others, altitude, heading,

airspeed, pitch, and power. Each set of values on these variables is called a state of the system.

A state is the complete set of variables needed to describe the entire performance of the pilot and the aircraft at any instant in time (Etkin, 1972). A state, as defined here, will contain more variables than needed by a pilot, an evaluator, or a training researcher. One objective of a control theory analysis is to select those items in a set of state variables that are necessary and sufficient to describe a standard flight path from start to finish. Brecke and Gerlach (1972) developed a model which uses descriptions of maneuvers given in the UPT syllabus as a basis for selecting the minimum set of variables and their values.

In time, a flight path is continuous and actually consists of an unlimited number of instantaneous states. Practically, however, the number of states must be limited by the number of discrete observations that can be made per specified unit of time. Pragmatically, observations taken at extremely high rates on a large number of variables result in quantities of data that are difficult to manage and process with a computer. The concept of a state can be extended using the concept of precision to obtain a solution to this problem of excessive detail.

## Steady States and Transition States

Many observations in a series on a system variable will be approximately the same value. A series of observations with approximately the same values can be regarded as a set of samples from a single

state. This extension of the concept of "state" is equivalent to extending the time frame from an instant to seconds, minutes, or longer periods. An extended state composed of contiguous samples with approximately equal values is called a steady state. It follows that length of time is a variable to include in the definition of extended states.

Few maneuvers in the UPT syllabus are purely steady states. All missions and most training maneuvers will consist of a sequence of steady states. The pilot must execute a sequence of changes in the aircraft controls to go from one steady state to the next. The series of instantaneous states in each period of change in control settings are grouped in an extended set called the transition state. Transition states are much more difficult to define and to work with in evaluating pilot performances (Connelly, Schuler, & Knoop, 1969; Dickman, 1974).[1]

Transition states are of special concern in evaluating student pilot performances for two reasons. First, new behaviors to be learned are usually located in the transition states (Gerlach, et al., 1972). For example, in the transition from climb to descent in the Vertical S-A, the student pilot must learn to coordinate smooth, gradual changes in pitch and power controls (see Figure 2). Second, the frequency of control changes specified in the mission objectives will be greatest at the transitions. Therefore, the probability of errors in changing the controls will be at a maximum in the vicinity of transition states.

---

[1]Some researchers, e.g., Leshowitz and Neilsen (1974, 1975), investigating the problems of evaluation in UPT have chosen not to work with the transition states for these reasons.

## Control Errors and Precision of Control

How are control errors and precision of control related? If the preceding analysis is correct, the largest variabilities in flight path data will be found in the vicinity of the transitions. For such conditions, the variance, a statistical measure of variability, is time dependent (Jacobs, 1969). Figure 3 depicts a set of time dependent observations on altitude and Figure 4 illustrates a time independent set. In Figure 3 each of the altitude curves exhibits a similar pattern of variation from the standard in time. The curves in Figure 4 reveal no discernible common pattern.

To the extent that similar control errors occur at a given transition and lead to similar deviations from the standard flight path, knowledge of any dependencies between time and variability would be useful. Methods need to be developed to infer the most probable control errors from objective data and to guide instructor pilot observations. Information about time dependencies could also be used diagnostically by instructional designers and instructor pilots to improve training. Researchers working to develop automated evaluation systems need this information to accurately specify criteria for processing data in the vicinity of the transitions (Dickman, 1974).

## Indicators Used in Pilot Training

Little is known about the specific control behaviors of pilots in the aircraft (Reid & Etkin, 1972). Existing models of pilot behavior are based on rigidly prescribed tracking tasks in the laboratory. Some practical knowledge about these behaviors is reflected

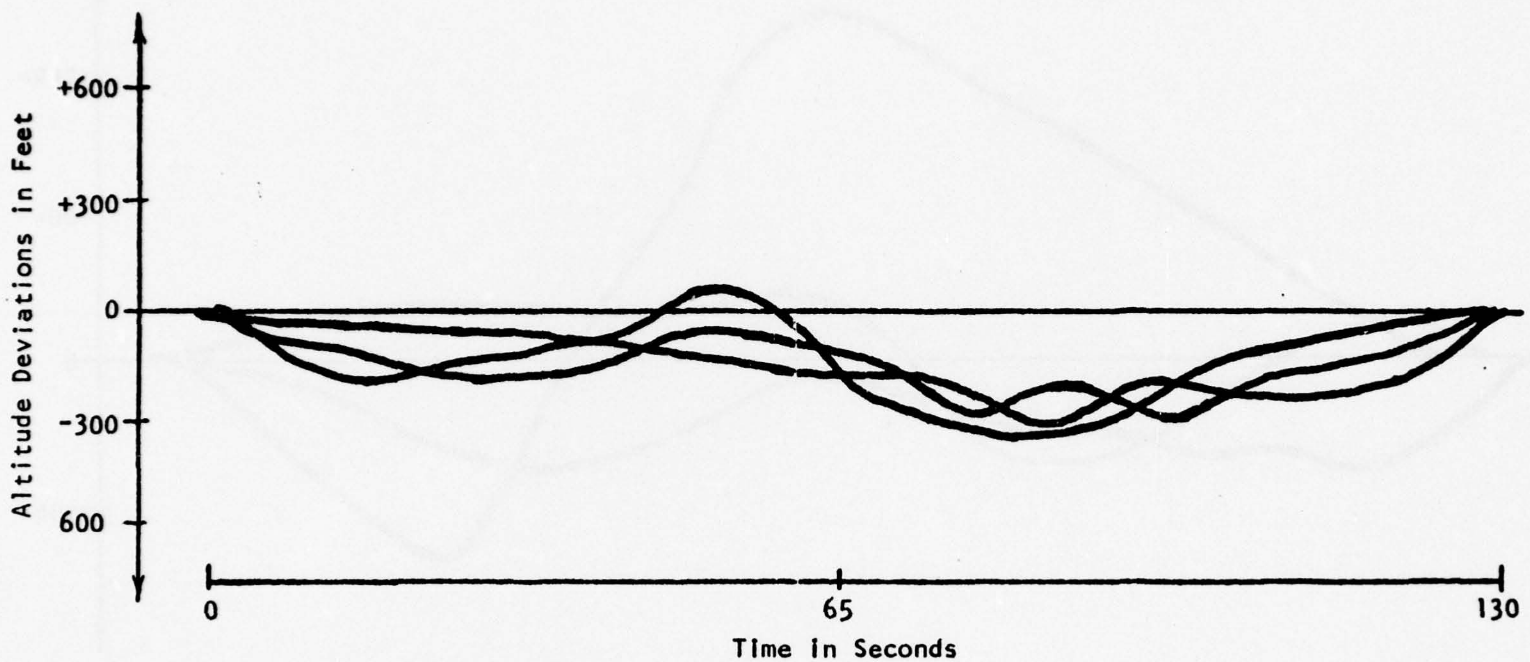


Figure 3.--Time Dependent Deviations From Vertical S-A Altitude Curve

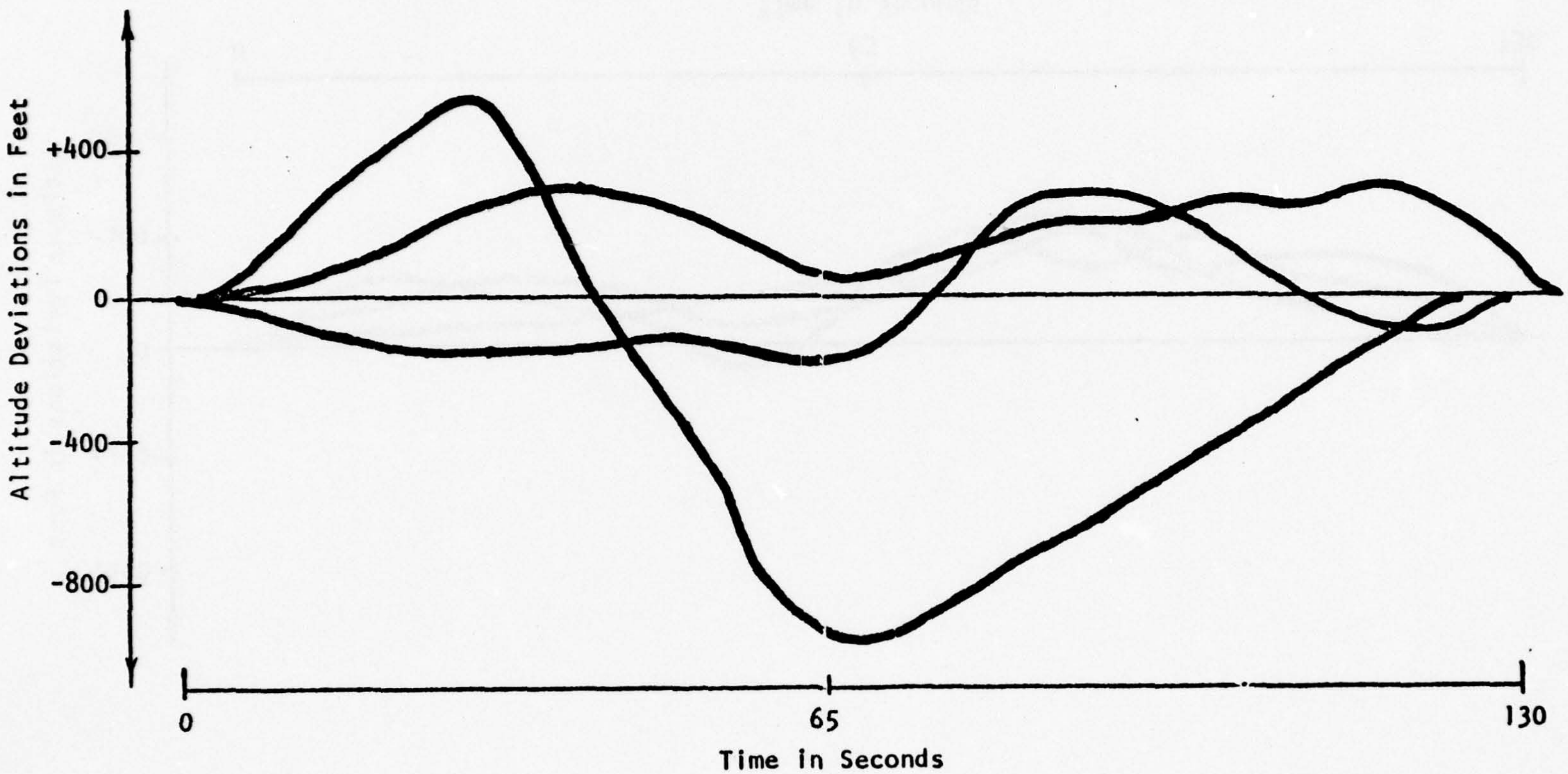


Figure 4.--Time Independent Deviations From Vertical S-A Altitude Curve

in the accumulated experiences of instructor pilots and is communicated as flight line lore (Reiser, Brecke, & Gerlach, 1972). However, instructor pilots are likely to disagree about proper control techniques and about how to interpret deviations from a standard flight path (Horner, Radinsky, & Fitzpatrick, 1970; Knoop & Welde, 1973; Reiser, et al., 1972). In the development of automated evaluation systems, questions about the meaning of deviations from the standard flight path must be solved (Connelly, et al., 1969; Dickman, 1974; Hill & Goebel, 1971; Knoop & Welde, 1973; Shipley, Gerlach, & Brecke, 1974).

Subjective Indicators

Rating scales are the predominant method of measuring pilot performance in the aircraft. A four-point scale--excellent, good, fair, and unsatisfactory--is used in undergraduate pilot training (UPT). The instructor pilot rates each maneuver during a student's training flight and he also gives an overall rating for the entire flight. Miller (1947) reviewed rating scales and paper and pencil observation schedules that were researched in the World War II pilot training studies. He concluded that improvements in evaluation methods would not be accomplished until automated data collection systems (ADCS) were available.

Instructor pilot ratings are not likely to be helpful in solving the problem of deviations between an observed and a standard flight path. Horner, et al. (1970) investigated the effects of differences between video taped performances on instructor pilot ratings. As student or instructor pilots performed specified UPT maneuvers in the aircraft, a video camera inside the cockpit was used to record views

of ground reference points and of the instrument panel. A set of these taped performances was selected to represent a range of variations from a video tape representing the standard flight path. Instructor pilots then rated these taped performances on a ten-point scale of performance quality. The ten-point scales were an expansion of the regular UPT four-point rating scale and the instructors used these expanded scales to rate segments of each maneuver. Horner, et al. found that variability among the ratings increased with the extent of the deviations between the rated tapes and the tape of the standard flight path. That is, as the extent of the deviations between tapes increased, the variability in ratings increased. Horner, et al. concluded that there are as many formulas for interpreting deviations as there are instructors.

Knoop and Welde (1973) had instructors and students perform a series of different UPT maneuvers in an aircraft. Measures of the system variables were recorded throughout each flight with an ADCS. At the completion of each maneuver during the flight, both the performer, a student or an instructor, and the accompanying instructor rated the performance. These ratings were made using the regular UPT four-point scale and each rating covered an entire maneuver. These subjective ratings were correlated with a number of different objective measures obtained from the ADCS recordings. On the basis of these correlations, Knoop and Welde concluded that instructor pilot ratings lacked standardization and that comparisons between ratings by different instructors would be unreliable.

Koonce (1974) used different methods and obtained large correlations ($\underline{r} \geq .80$) between pairs of raters. Instrument rated, highly qualified pilots were tested with a series of maneuvers for a commercial pilot's instrument rating. Certified flight instructors were used in pairs to observe each performance and they recorded data using a paper and pencil observation schedule. The observers were briefed on the use of the observation schedule, given training on the observation task, and experience in performing the task in a flight simulator. No objective criteria were used in training the observers; when an observer reported that he was ready, he was assigned to the roster of observers. Each observer was paired with every other observer in a counter-balanced design and each pair collected data in both the simulator and the aircraft.

The studies reviewed in this section lead to two conclusions. First, instructor pilots must be trained to give standard evaluations of variability. Second, criteria must be developed so that deviations between a standard flight path and observed data can be consistently interpreted. It follows that standard criteria must be developed before instructors can be trained to give standard evaluations. If comparisons between simulator and aircraft training are to involve the use of instructor ratings, similar criteria must be used in each training situation. Practically, advanced simulation systems can be used to develop standard criteria and to train the instructor pilots to use these criteria.

Objective Indicators

One problem is to obtain objective indicators of variability in pilot performance data which can be used as criteria to evaluate precision of control. In studies reviewed on this problem, the indicators were computed from objective data with automated systems. To obtain objective data from pilot performances in an aircraft or in a flight simulator, special electronic, ADCS, devices are connected to selected system output variables (Hill & Goebel, 1971; Knoop & Welde, 1973; Shipley, et al., 1974). These ADCS devices sample and compute summary measures or record data points from each output variable at small, fixed intervals of time, e.g., one discrete data point per variable per second. Data obtained with these devices form a time series and, for the purposes of evaluation with pilot performances, conventional indicators of variability may not be adequate (Winer, 1971).

A standard deviation is a conventional measure of precision in a set of data. Hill and Goebel (1971) found some weak evidence to support the use of the standard deviation as a measure of performance skill in pilot training. They computed standard deviations and other measures with a small on-line computer attached to a GAT-1 simulator. The measures were computed from performance variables that were assigned constant values in the testing tasks, i.e., variables with changing values were not sampled. Methodologically, 266 different statistical values were computed and tested with analysis of variance for the differences among three performance groups at the .05 level. Forty tests were significant (.05) and 13 of these were standard deviations.

There is reason to suspect the use of a sample standard deviation and similar estimators of precision of control when computed across a set of time series data. The problem arises if a sample measure of central tendency is used as the basis of the measure of deviations. If the data is sampled from a nonlinear variable and the sample mean is used as the measure of central tendency, the measures of precision will be too large, thus underestimating the real level of skill. A second case of inaccurate estimation occurs if all sampled observations deviate in the same direction relative to an assumed standard, e.g., all positive. In this latter case, estimators of precision calculated from the sample mean will be too small, thus overestimating precision. Clearly, estimates of precision of control must involve a reference to some independent standard.

In describing relationships between training objectives and requirements for evaluation in pilot training, Brecke and Gerlach (1972) consider the use of performance limits:

> It is impossible for either the IP or the SP to make more than a very rough subjective evaluation of pilot performance unless some objective criterion limits are prescribed which, at the very least, permit a distinction between acceptable and unacceptable performance. Ideally, however, such performance or criterion limits should clearly and objectively define ranges of performance in accordance with the grading and evaluation system currently in use. (p. 10)

A performance limit is a value that can be used to differentiate between acceptable and unacceptable performances (Figure 5). As an indicator of acceptable performance, a performance limit expresses the maximum allowable variability from the standard. Percent time on criterion is

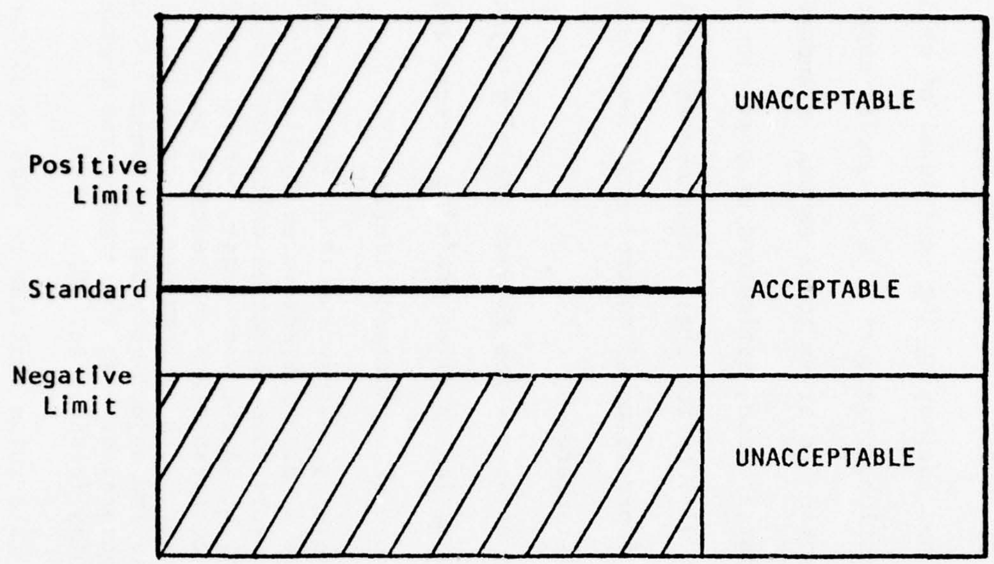


Figure 5.--Performance Limits in a Dichotomous Evaluation Model

one indicator of central tendency used with performance limits (Fitts, Bahrick, Briggs, & Noble, 1959).

Percent time indicators should not be used exclusively to evaluate pilot performances (Fitts, et al., 1959). Three weaknesses of dichotomous performance limit indicators have been identified: (a) difficulty in specifying the size of the limits; (b) sensitivity of limits to changes in variability; and (c) need to evaluate effects of training on errors in performance data. There are difficulties in specifying the size of the performance limits. Horner, et al. (1970) and Knoop and Welde (1973) have shown that instructor pilot judgments are not based on reliable standards and should not be used exclusively to determine the size of performance limits. Knoop and Welde recommend that within subject sampling procedures be used as the basis for defining objective measures. By within subject sampling, Knoop and Welde mean that each pilot performs a series of trials on the assigned maneuver. With such performances, Fitts, et al. (1959) report that the standard deviation, measured from the group mean, will give maximum discriminations between different performances.

Relative to the standard, estimates of learning effects will be sensitive to differences in the size of the performance limits (Fitts, et al., 1959). Early in training, if the limits are too small, large changes in variability will not be detected. Near the end of training, small changes in variability will not be detected if the limits are too large. To the extent that changes in variability reflect improvements in precision of control, estimates of learning based exclusively on percent time indicators will not be valid in pilot training.

Potential interactions caused by changes in the performance limits are seen in data abstracted from Shipley, et al. (1974). Fitts, et al. (1959) described a procedure for estimating percent time on criterion from the normal probability distribution. Shipley, et al. used these procedures to obtain percent time estimates with different performance limits from the same set of data. Evidence for potential interaction effects can be seen in these results (Table 1).

In a simple dichotomous evaluation model, there is no way to evaluate the effects of training on the unacceptable parts of a performance. The concept of error amplitude can be used to evaluate these data. The concept of error amplitude is illustrated in Figure 6. Mathematically, error amplitude is expressed as a ratio of the deviation between an error observation exceeding the nearer performance limit to the size of the limit. Statistically, error amplitude is summarized over an entire performance as a root mean square (RMS). The mean square is computed from the total number of observations in the data, T, or from the number of deviations, D. For the data from the example in Figure 6, the error amplitudes, E(T) and E(D) are: T = 13; D = 4; and

$$E(T) = \sqrt{\frac{1.1^2 + 2.8^2 + 1.4^2 + (-0.4^2)}{13}} = .93; \text{ and,}$$

$$E(D) = \sqrt{\frac{1.1^2 + 2.8^2 + 1.4^2 + (-0.4^2)}{4}} = 1.67.$$

TABLE 1

CHANGES IN TWO STATISTICAL MEASURES CAUSED BY DIFFERENT CRITERION LIMITS[a]

| Performance Group | Percent Time on Heading: Standard Deviations Criterion Limits | | |
|---|---|---|---|
| | 1 Degree (A) | 5 Degrees (B) | Proportional Increase (B-A)/A |
| 1 | 6.45 | 21.44 | 2.32 |
| 2 | 5.17 | 18.33 | 2.55 |
| 3 | 5.29 | 17.54 | 2.32 |

| Performance Group | Percent Time on Altitude: Means Criterion Limits | | |
|---|---|---|---|
| | 50 Feet (A) | 100 Feet (B) | Proportional Increase (B-A)/A |
| 1 | 61.87 | 79.98 | .29 |
| 2 | 40.85 | 68.94 | .69 |
| 3 | 36.31 | 52.06 | .43 |

[a]Data taken from Table 3 (p. 33) and Table 6 (p. 36), Shipley, _et al_., 1974.

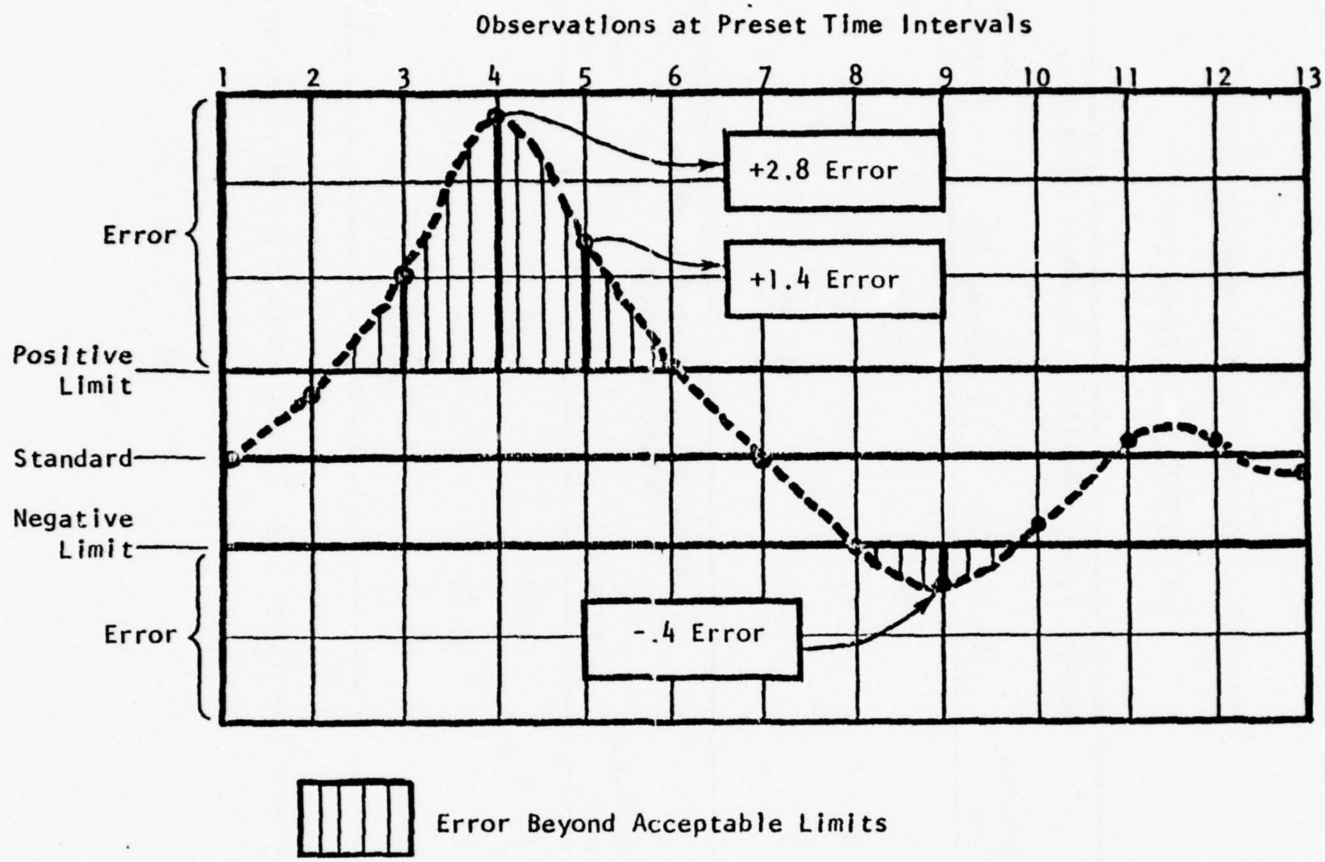


Figure 6.--A Performance Limits Model for Error Evaluation

Conceptually, the measure E(T) indicates the average rate of error magnitude over all observations and E(D) indicates the average size of each error. Each of these error amplitude indicators is an estimate of the precision of control in a set of data.

Shipley, et al. (1974) combined error amplitude, E(T), with an indicator of time on criterion, called hit rate, to evaluate student pilot performances in a training experiment. Hit rate is the number of observations within the performance limits divided by the total number of observations. Brecke (1975) and Brecke, Gerlach, and Shipley (1974) evaluated the effects of preflight instructions on student performances in a flight simulator with hit rate, error amplitude, and the Fitts, et al. (1959) percent time on criterion indicators. In the Brecke, et al. (1974) study, there were no significant effects on any performance variable with any of these indicators. In the Brecke (1975) study, both error amplitude, E(T), and Fitts, et al. percent time indicated a significant interaction between type of instruction and performance trials. Hit rate showed no effects in either study. The source of the difference was found in rate of improvement in precision of control across trials. Students given the experimental instructions performed consistently well across all learning trials. Those given regular instructions showed definite improvements in precision of control on the first two trials.

## Alternative Indicators

Performance limit indicators are effective to a limited extent in the comparisons of group performances on entire trials. These

summary indicators obliterate details and become increasingly insensitive to specific deviations as the number of observations increases. Indicators that are more sensitive to differences in performance states might be even more effective. Maximum deviation and performance time are two indicators that can be used in place of those based on performance limits.

A review of indicators used in previous pilot training research produced three possibilities which might be used in a preliminary, then a detailed analysis: the range, maximum deviation from the standard, and performance time. Hagin (1947) established the range and maximum deviation as objective indicators in the evaluation of instrument flying skills. In parallel research during World War II, performance time was also studied (Miller, 1947). Each of these indicators is considered in the context of the present research. In the present research, the emphasis is on the use of comprehensive data collected and evaluated by automated devices. In the previous research, instructor and check pilots had to record the data using paper and pencil.

Miller (1947) summarized the results of measurement research in pilot training carried out in the World War II military studies. Three methods of scoring were compared: (a) time sampling of deviations at specified intervals; (b) the range; and (c) maximum deviation. As compared to instructor ratings, each of the methods was equally valid. Time sampled deviations were more reliable than range and maximum deviation on trials in the same performance. These deviations were generally equivalent to the range and maximum deviation on test-retest reliabilities across days and observers. However, at that time it was

difficult to use time sampling procedures and limited evidence suggested that the range might be more reliable on test-retest than the maximum deviation.

The deviations method can be used with ADCS devices because these devices collect data with fixed interval time sampling methods. If the objective is to evaluate variability relative to a standard, deviations are preferable to the range on logical grounds. Maximum deviation will be a better indicator of skill in maintaining a standard than the range under certain conditions. These conditions occur if the two values defining the range are located (a) about equally distant but in opposite directions from the standard, or (b) at some distance in the same direction from the standard. For two values located in opposite directions from the standard, the range may be as much as two times larger than the maximum deviation. For values located in the same direction from the standard, the maximum deviation will equal or exceed the range (see Figure 3, page 18).

In time series data with the conditions just described, maximum deviation will be a better indicator of precision than the range because maximum and minimum observations are not independent. There is no possibility of finding these two values adjacent to each other in any small time interval, say several seconds. In such data, a maximum deviation and its direction from the standard is more informative than the range because it tells us that deviations in some related small time interval will be similar in direction and magnitude. Therefore, a maximum deviation and its time of occurrence will locate a pattern of deviations in a performance; the range will not.

Performance time is not as well established as an indicator of performance quality in pilot training as either the range or maximum deviation. In the World War II studies, performance time was found to be a good indicator for some maneuvers but not for others (Miller, 1947). As a difficult measure for pilots to observe, it was dropped from the early observation schedules. Some recent evidence from data collected with ADCS in the aircraft (Knoop & Welde, 1973) and in a simulator (Shipley, et al., 1974) supports a need to further investigate performance time.

Performance time was found to be a good indicator in two recent studies. In a training experiment, Brecke, et al. (1974) found that differences in means and variances for performance time differentiated between treatment groups. The Brecke, et al. data was obtained automatically from student pilot performances in a flight simulator (Shipley, et al., 1974). Knoop and Welde (1973) obtained 87 different objective indicators from data with ADCS methods in an aircraft. These data included performances by senior instructor pilots, instructor pilots, and student pilots. The 87 objective indicators were correlated with two different ratings of each performance, one by the performer and one by the instructor. Of the 87 different indicators, 7 or 8% were time measures; of the 77 significant, .05, correlations obtained, time measures accounted for 11 or 14%.

In the World War II studies, performance time was found to be a significant indicator of performance quality if maneuver objectives prescribed starting and ending points and the form of the desired flight path between them, e.g., a maximum performance turn (Miller,

1947). Other maneuvers and maneuver states in the pilot training curriculum possess similar characteristics. In some cases, the desired flight path is not clearly specified in existing materials, but it can be inferred and checked empirically (Brecke & Gerlach, 1972; Knoop & Welde, 1973).

To the extent that performance time can be empirically validated in a given maneuver, it is a potentially powerful indicator of performance characteristics. Performance time is related to the aircraft's flight path through the laws of motion and aerodynamics (Etkin, 1972). It is logically possible for an observed flight path to meet the requirement of a standard time and still contain important deviations from the standard flight path. Such conditions will occur in case of large compensating errors at different locations in a performance. However, it is impossible for a performance to deviate from the standard time and also be on all criteria throughout the entire flight.

## A Performance State Evaluation Model

Performance time can be used with maximum deviation to simplify performance evaluations for selected maneuvers. To accomplish this objective, one designates states of a performance and the flight path conditions which must be satisfied at the endpoints of each state (Brecke & Gerlach, 1972). Time is a variable in the set defining each state. If the existing materials do not specify desired time values, they can be inferred or obtained empirically from mastery performances of experienced pilots. Determining the level of experience needed for reliable estimates must be investigated.

An algorithmic procedure is used to apply performance time with existing ADCS data (Figure 7). If total time fails, some state or states must be in error. The state or states in error can be located by examining performance time for each state. For a state that fails the time test, if the entry conditions to that state fail the deviations test, at least some source of the error will be in the preceding state or states. If the entry conditions for a state are satisfied, i.e., all values are within the performance limits, the source of the error must be located in some set of deviations between the entry point and the end point of that state. If the end point conditions fail, the next state should also exhibit some deviations.

Differences in means and variances for performance time were strongly associated with similar differences in maximum altitude for student performances on the Vertical S-A (Brecke, et al., 1974). Maximum altitude is the endpoint between two adjacent transition states in the Vertical S-A. Gerlach, Brecke, Reiser, and Shipley (1972) predicted that students would have their greatest difficulty in mastering the Vertical S-A performance in these two transition states. Further analysis of these data revealed that the statistical differences on these measures were related to differences in specific instructions about pitch control (Brecke, et al., 1974).

In a subsequent study, Brecke (1975) found significant interactions on error amplitude between type of instruction and performance trial. Percent time and hit rate did not reveal similar differences. Comparisons between treatment groups and a control group were not

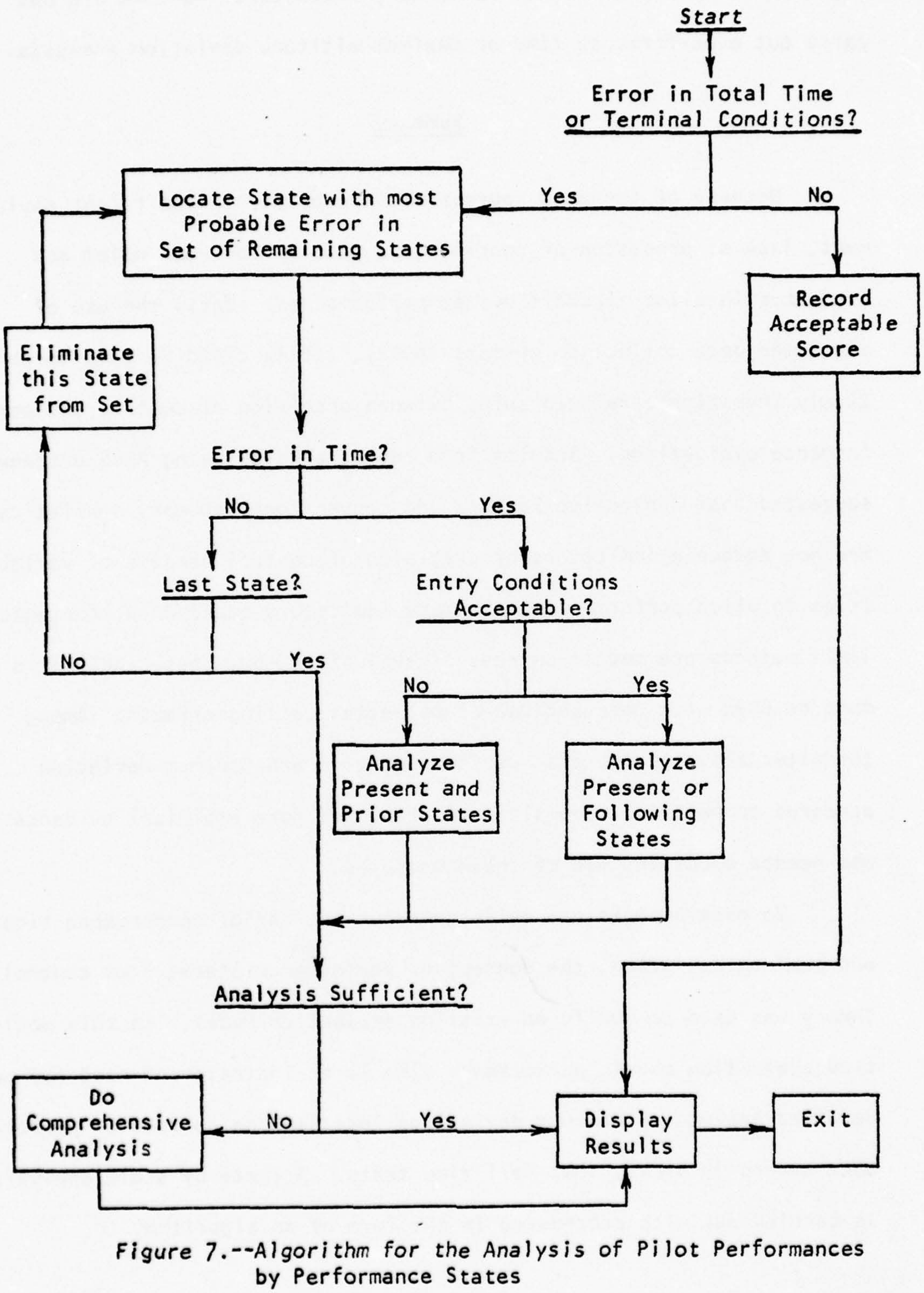


*Figure 7.--Algorithm for the Analysis of Pilot Performances by Performance States*

significant on any of the three summary indicators. Brecke did not carry out a performance time or maximum altitude deviation analysis.

## Summary

Because of the many sources of variability in the flight environment, lack of precision of control is a problem for both pilot and evaluator in pilot-aircraft system performances. Until the use of automated data collection systems (ADCS), little could be done to objectively investigate relationships between precision of control and performance evaluations. Results from recent studies using ADCS procedures suggested that subjective ratings and conventional summary statistics are not adequate indicators of precision of control because of variabilities in pilot performances. Summary indicators based on performance limit methods are better sources of evaluation, but these indicators must be used with care because of potential ceiling effects. Among the alternative indicators, performance time and maximum deviation appeared to meet the requisite criteria, but more empirical evidence was needed about the use of these measures.

To obtain empirical evidence about the use of performance time and maximum deviation, the concept of performance states from control theory was used to modify an existing evaluation model. In this modified evaluation model, performance time is an indicator of need for more detailed analyses. Maximum deviations indicate the sources of performance errors in states that fail time tests. A state by state analysis is carried out with procedures in the form of an algorithm.

# CHAPTER III

## METHODS

Three empirical investigations were carried out with objective data to test a performance state evaluation model. In this model, performance times from objective data are used as preliminary indicators of performance quality and deviations from standard values are used to identify specific errors. The objective data were obtained from pilot performances in a flight simulator with ADCS (Shipley, Gerlach, & Brecke, 1974). Data from performances by 2 experienced and 39 student pilots on an instrument flight maneuver were used in these investigations. In the first investigation, it was hypothesized that there would be no significant differences between the performances of two experienced pilots on performance times compared with each other and with standard time values. In the second investigation, total performance times were analyzed with analysis of variance using a 2 x 2 mixed design and Dunnett's procedures were used to compare each treatment group to an external control group. Maximum altitude variances were tested for differences using F-ratios. Data for the second investigation were taken from an experimental study by Brecke (1975).

The first two investigations were carried out to establish a basis to design the third and final investigation. In the final investigation, the hypothesis was that performance times, maximum altitude, and personal data about each subject would predict error amplitude

scores. The error amplitude scores from the Brecke (1975) study were used as the criterion variable in a stepwise regression analysis. The variates in the regression analysis were transformed time and maximum altitude values and personal data for each subject. In this chapter, the general methods used in the three investigations are described, the first two investigations and their results are reported, and the design and procedures used in the third investigation are presented.

## A Standard Flight Path

### A Test Performance

An instrument flight maneuver, the Vertical S-A, was used as the basis of the performances examined in the three investigations. Brecke and Gerlach (1972) analyzed materials from the Air Force Undergraduate Pilot Training syllabus dealing with the Vertical S-A. They developed a comprehensive set of values to represent the flight path for the purposes of instructional development (Brecke, 1975; Brecke, Gerlach, & Shipley, 1974; Gerlach, Brecke, Reiser, & Shipley, 1972). Shipley, Gerlach, and Brecke (1974) also used these values and training objectives to develop ADCS methods to evaluate performances of student pilots participating in training experiments. In the present investigations, the standard Vertical S-A flight path consisted of a sequence of seven states: six extended states and one momentary state. The seven states in the standard Vertical S-A flight path are:

1. the transition from level flight to climbing flight;
2. climbing flight;
3. the transition from climbing flight to maximum altitude;

4. maximum altitude;

5. the transition from maximum altitude to descending flight;

6. descending flight; and

7. the transition from descending flight to level flight.

(These states are illustrated on the standard Vertical S-A altitude curve in Figure 2, page 12.)

Expected Times

The flight path for the Vertical S-A, as described in the training objectives, must be symmetric about the maximum altitude. That is, the values for States 1 to 3 are identical to the values for States 5 to 7 except for the difference in direction: climb versus descent. The principle of symmetry was used to derive the time to perform each state in the standard flight path. Each state covers a specified change of altitude; for example, 800 feet each in States 2 and 6. The rate of vertical change is also specified at 1,000 feet per minute for these two states. It follows that the time to perform these 2 states is 48 seconds each (800 feet divided by 16.67 feet per second).

In the transitions, empirical estimates of times were necessary because the vertical rate is changing. Estimates of the time to perform the transition from climb to descent, States 3 to 5, were obtained from observations of an experienced pilot's performance. An instructor pilot from the Air Force Human Resources Laboratory, Flight Training Division, at Williams Air Force Base performed a series of ten Vertical S-A maneuvers in a flight simulator. The average performance time from 15,900

to 16,000 and back to 15,900 feet was 16.38 seconds with a standard deviation of 1.78 seconds.

Since the maneuver is symmetrical, it was assumed that each transition state would require an equal performance time. The training objectives specify that each transition covers 100 feet of altitude change and that each has similar starting or ending rates of vertical change: to or from 0 feet per second and to or from plus or minus 16.67 feet per second. An estimate of about 8 seconds (16.38 seconds divided by 2) was obtained as the standard time for each of the transition states. The total time for the standard Vertical S-A flight path used in these studies was 129 seconds: 96 seconds for the 2 steady states, 32 seconds for the 4 transition states, and 1 second for maximum altitude. The standard time, altitude, and average vertical rate values for each state are summarized in Table 2.

## Data Collection Procedures

In the three investigations carried out in this study, data were obtained from an experimental study by Brecke (1975). In Brecke's experiment, measures of time, airspeed, heading, vertical rate, pitch, power, and altitude were collected. With ADCS set at an interval of one second, data on these measures were collected from performances of student pilots in a flight simulator. In the present study, elapsed time values at the endpoints of each of the seven states in the Vertical S-A, $T_1$ to $T_8$, and values for starting, maximum, and ending altitude, $A_1$ to $A_3$, were obtained from the data recorded by ADCS. Procedures used to obtain these values are summarized in Table 3.

TABLE 2

STANDARD TIMES, ALTITUDE CHANGES, AND VERTICAL RATES

FOR EACH PERFORMANCE STATE IN THE VERTICAL S-A

| State | Time | Altitude Change | Average Vertical Rate |
|---|---|---|---|
| 1 | 8.00 sec | +100 ft | +12.50 ft/sec |
| 2 | 48.00 sec | +800 ft | +16.67 ft/sec |
| 3 | 8.00 sec | +100 ft | +12.50 ft/sec |
| 4 | 1.00 sec | 0 ft | 0 ft/sec |
| 5 | 8.00 sec | -100 ft | -12.50 ft/sec |
| 6 | 48.00 sec | -800 ft | -16.67 ft/sec |
| 7 | 8.00 sec | -100 ft | -12.50 ft/sec |
| TOTAL | 129.00 sec | 2000 ft | 15.50 ft/sec |

TABLE 3

DATA COLLECTION SCHEDULE

| Event | Time[a] | Altitude | Comments |
|---|---|---|---|
| Start | $T_1$ | $A_1$ | |
| 15,100 feet | $T_2$ | | If starting altitude is greater than 15,100 feet, $T_2 = T_1$. |
| 15,900 feet | $T_3$ | | If maximum altitude occurs before 15,900 feet, $T_3 = T_4$. |
| Maximum altitude | $T_4$, $T_5$ | $A_2$ | If maximum altitude occurs once, $T_5 = T_4$. If maximum altitude occurs more than once, $T_5$ = time at last occurrence. |
| 15,900 feet | $T_6$ | | If maximum altitude is less than 15,900 feet, $T_6 = T_5$. |
| 15,100 feet | $T_7$ | | If end occurs before 15,100 feet, $T_7 = T_8$. |
| End | $T_8$ | $A_3$ | |

[a]Total time (TOTTIM) and maximum altitude time were defined inclusively, e.g., TOTTIM = $T_8 - T_1 + 1$; all other performance state times were simple differences, e.g., $T_2 - T_1$.

The conditions of performance and the characteristics of the ADCS data did not always permit altitude observations that were precisely 15,100 or 15,900 feet. In such cases, the nearest altitude values to these criterion points were used. In determining the starting and ending points, altitude values were used as the primary indicator. In all cases the altitude value of 15,000 feet or the value nearest to 15,000 feet was used to indicate the start and end of the trial. For all 234 trials in the Brecke (1975) study, the mean (M) and standard deviation (SD) for starting and ending altitudes were (a) starting: M = 14,980 feet, SD = 206 feet; and (b) ending: M = 14,940 feet, SD = 320 feet.

Some cases were identified where pitch and power values indicated that a trial had started or ended at an altitude other than 15,000 feet. The typical range for starting altitude was from 14,800 to 15,200 feet. In cases of ADCS recording malfunctions, starting or ending altitudes were observed at 15,700 feet, but adjacent altitude values were also observed with differences as great as 600 feet. In such cases, the pattern of altitude values, and values from power and pitch, were used to identify the point in question and to derive its actual value.

## Preliminary Investigation I

Investigation I was carried out to compare the standard time values in Table 2 with the performance times of two experienced pilots and to discover whether or not the performance state evaluation model would reveal any differences between the performances of these two pilots. If the performance state evaluation model detected differences between performance times for experienced individual pilots, it would also

discriminate between performances of student pilots. Discriminations between performances and standard values and between different performances were considered essential to effective evaluations in regular training and in experimental research on the effects of training methods.

Method

Two experienced pilots, one a researcher and the other an instructor pilot, each performed a sequence of six trials on the Vertical S-A maneuver in a flight simulator according to procedures described by Brecke (1975). Originally, these performances were obtained as a part of the tryout procedures of the ADCS device. Trials 1 and 6 of the instructor pilot's performances were not included in the present investigations. Trial 1 was a descending rather than a climbing Vertical S-A; Trial 6 was unusable because of an ADCS malfunction.

Procedures

Performance time and maximum altitude data were obtained from computer printouts of the ten remaining performances (see Table 3). Means and standard deviations for total time, time for each performance state, and maximum altitude were computed for each pilot's data. The means were tested with *t*-tests at the .05 level. Absolute values were obtained for deviations of observations from corresponding standard values on each of the indicators named above. Standard deviations, $\sigma$, were computed from these absolute deviations for each indicator to use in subsequent research. Tests were also carried out with these deviation values using normal deviate, *z*-score, methods to identify

performances or performance states that were significantly different from the standard values at the .05 level. Each statistical test was made using two-tailed values and the tests on the standard values were made using confidence limit procedures (Hays, 1963).

Results

The means, standard deviations, and 95% confidence limits for each indicator are given in Table 4. Three significant differences were found in the sample values. These significant differences were the means for total times: $\underline{t} = 4.02$, df = 8; the means for State 2 times: $\underline{t} = 2.64$, df = 8; and the variances for State 5 times: not homogeneous, $F_{max} = 20.95$ on 1 and 3 degrees of freedom (significant at or beyond the .05 level using Hartley's test [Myers, 1966]).

Among the *total time* deviations, one performance by the researcher pilot was found to be significantly different. The total time deviation for his third performance was 24 seconds less than expected. Table 4 includes a list of 95% confidence limits for deviations from standard times. On the basis of these limits, one performance state by the researcher pilot was also identified as significantly deviant. In performance one the time for State 6 was 15 seconds less than expected. No performance times of the instructor pilot were significantly deviant.

There were no significant differences between maximum altitudes for these two experienced pilots. Means and standard deviations for maximum altitude were: (a) for the researcher pilot: M = 15,987.78 feet, SD = 27.37 feet; and (b) for the instructor pilot: M = 15,995.36 feet, SD = 27.27 feet.

TABLE 4

MEANS, STANDARD DEVIATIONS, AVERAGE ABSOLUTE DEVIATIONS (ABS), AND 95% CONFIDENCE LIMITS ON PERFORMANCE TIMES FROM TEN VERTICAL S-A PERFORMANCES BY TWO EXPERIENCED PILOTS

| Performance State | Researcher Pilot | | | Instructor Pilot | | | | 95% |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | ABS | Mean | SD | ABS | σ | Limits[a] |
| 1 | 7.83 | 1.60 | 1.17 | 8.25 | 2.06 | 1.25 | 1.61 | ± 3.16 |
| 2 | 43.67 | 6.25 | 5.67 | 53.75 | 5.32 | 5.75 | 7.25 | ± 14.21 |
| 3 | 6.83 | 2.99 | 2.50 | 7.75 | 2.99 | 2.25 | 2.83 | ± 5.55 |
| 4 | 2.33 | 1.51 | 1.33 | 2.00 | 1.41 | 1.00 | 1.79 | ± 3.51 |
| 5 | 3.33 | .52 | 4.67 | 4.50 | 2.38 | 3.50 | 4.77 | ± 9.35 |
| 6 | 44.17 | 6.71 | 5.17 | 45.75 | 4.27 | 3.75 | 6.23 | ± 12.21 |
| 7 | 7.00 | 2.19 | 2.00 | 7.50 | 1.29 | 1.00 | 1.90 | ± 3.72 |
| TOTAL | 115.15 | 5.98 | 13.83 | 129.50 | 4.65 | 3.50 | 11.80 | ± 23.13 |

[a]These limits are for use with deviations of an observed from a standard time in seconds.

Discussion

It was possible to discriminate between the average performances of two experienced pilots on the basis of their performance times. Likewise, it was possible to locate significant deviations in their performance states by attending to deviations from the standard times. Acceptable total performance times must be interpreted carefully because patterns of too little and too much time in the performance states will cancel each other. For example, in the first performance of the researcher pilot, a total time deviation of -9 seconds was not significant, although the deviation of -15 seconds for State 6 was significant. In this performance, accumulated deviation time for performance States 1 to 3 was +11 seconds, while the accumulated deviation time for performance States 5 and 6 was -20 seconds. On the other hand, a significant total deviation time may be accumulated over a series of performance states. In performance three by the researcher pilot, the total deviation, -24 seconds, was accumulated as -28 seconds across States 1, 2, 5, 6, and 7 although none of the deviations for an individual state was significant.

The statistical tests carried out in the present investigation were based on an assumption that performance times were normally distributed about the standard value of 129 seconds. That assumption is questionable for the data in the ten performances examined in this investigation. The performances of the instructor pilot ranged from 124-135 seconds; the mean time of his four trials, 129.5, was equivalent to the standard total time of 129 seconds. However, for the researcher pilot all the performance times ranged from 105-122 seconds, and his mean total

time, 115.15, was significantly less ($\underline{t}$ = -5.18, df = 5, $\underline{p} < .01$) than the standard time of 129 seconds. A definitive test of the assumption of normally distributed time values cannot be made, however, without more data from performances of experienced pilots. The standard times from Table 2 and the estimated population standard deviations ($\sigma$) from Table 4 were used in the main investigation in this study.

## Preliminary Investigation II

Investigation II was carried out to determine whether total performance time or maximum altitude would discriminate between performances of treatment groups in a training experiment. An *a priori* prediction was that differences among maximum altitude variances would discriminate among treatment group performances. This prediction was based on a performance analysis (Gerlach, Brecke, Reiser, & Shipley, 1972) and on the outcomes of a prior experiment (Brecke, Gerlach, & Shipley, 1974). The values used in the present investigation were obtained from performance data produced by Brecke (1975) in an extension of the 1974 experimental research.

### Method

Thirty-nine student pilots were randomly assigned to one of five groups. Members of the four experimental groups studied objectives and different preflight instructions on how to perform the Vertical S-A; members of a control group studied only the objectives. After each subject had studied his assigned materials, he performed a sequence of six trials on the Vertical S-A according to procedures described by Brecke

(1975). Data on each subject's performances in the flight simulator was collected using ADCS.

Procedures

Total performance time and maximum altitude values were obtained from the Brecke data using procedures described in Table 3.

Performance time. The total time values were analyzed using analysis of variance with a two between- one within-subjects design; trials was the within subjects variable. Dunnett's method (Myers, 1966) was used to compare the means of the treatment groups to the means of the control group. The statistical hypotheses of no significant main effects or interactions and no significant contrasts between treatment and control groups were tested.

Maximum altitude. On the basis of a performance analysis, differences in performances were to be expected at the maximum altitude (Gerlach, et al., 1972). In previous research, differences in treatment group variances (heterogeneity of variance) were found to be significant indicators of treatment effects (Brecke, et al., 1974). Tests of differences in variances were used in this investigation rather than tests for differences in means, i.e., ANOVA or $t$-tests. The hypotheses were based on a priori predictions of expected differences in the variabilities of performances among the treatment groups.

In the present investigation, it was hypothesized that there would be significant differences, .05, among variances on maximum altitude given type of experimental treatment. There were five treatment groups (a 2 x 2 design with control group) in the Brecke (1975) study.

It was predicted that the variances for two groups ($A_1B_1$ and $A_1B_2$) receiving experimental instruction would be significantly smaller, .05, than variances for the control group (C) or variances for two groups ($A_2B_1$ and $A_2B_2$) receiving current instruction. A one-tailed test was used with this directional hypothesis and two-tailed tests were used to test the hypothesis of no other significant differences among the remaining group variances. The critical $\underline{F}$ values used to carry out these tests are given in Table 5.

## Results

Total time. In the total time ANOVA, there were significant main effects on type of instruction, number of practice items in the instructional program, and performance trials. A significant interaction was also observed between levels of practice items and performance trials. The ANOVA summary is given in Table 6 and the means for the significant interaction and trials are given in Table 7. The means for the effects due to type of instruction were: (a) experimental, 111.67, and (b) current, 131.16. For the effects due to levels of practice, the means were: (a) low, 128.10, and (b) high, 114.72. There were no significant contrasts on Dunnett's test between any treatment group and the control group.

Maximum altitude. The comparisons among the treatment group variances were significant as predicted (Table 8). At maximum altitude, variances for the two groups receiving experimental instructions were significantly (.05) smaller than the variance for the control group and than variances for the two groups receiving current instruction. The observed $\underline{F}$-ratios in these contrasts range from 4.06 to 10.39. The

TABLE 5

CRITICAL $\underline{F}$-VALUES FOR TESTS OF PREDICTED DIFFERENCES IN VARIANCES

| Experimental | | Treatment Control | Current | |
|---|---|---|---|---|
| $A_1B_1$ | $A_1B_2$ | C | $A_2B_1$ | $A_2B_2$ |
| -- | 1.88[a] | 1.69[b] | 1.69[b] | 1.69[b] |
| | -- | 1.69[b] | 1.69[b] | 1.69[b] |
| | | -- | 1.88[a] | 1.88[a] |
| | | | -- | 1.88[a] |

[a]Two-tailed tests at .05.

[b]One-tailed tests at .05.

NOTE: There were either 42 (Group C) or 48 observations in each variance. The critical values in this table are based on 40 degrees of freedom for each variance. These degrees of freedom lead to slightly conservative tests. The two-tailed values are taken from tables at .025.

TABLE 6

ANALYSIS OF VARIANCE: TOTAL TIME

| Source | SS | df | MS | F | p |
|---|---|---|---|---|---|
| Total | 157872.50 | 191 | 826.56 | | |
| Between | 73560.66 | 31 | 2372.92 | | |
| Instruction (A) | 18232.51 | 1 | 18232.51 | 10.95 | .005 |
| Practice Items (B) | 8600.13 | 1 | 8600.13 | 5.16 | .05 |
| A x B | 49.01 | 1 | 49.01 | .03 | n.s. |
| Error | 46679.02 | 28 | 1667.11 | | |
| Within | 84311.84 | 160 | 526.95 | | |
| Trials (T) | 9314.71 | 5 | 1862.94 | 4.20 | .005 |
| T x A | 3564.33 | 5 | 712.87 | 1.61 | n.s. |
| T x B | 5151.34 | 5 | 1030.27 | 2.32 | .05 |
| T x A x B | 4143.83 | 5 | 828.77 | 1.87 | n.s. |
| Error | 62137.62 | 140 | 443.84 | | |

TABLE 7

MEANS OF SIGNIFICANT EFFECTS OF TRIALS AND PRACTICE BY TRIALS INTERACTION ON PERFORMANCE TIME[a]

| Practice Items | Trials | | | | | |
|---|---|---|---|---|---|---|
| | One | Two | Three | Four | Five | Six |
| Low | 146.69 | 115.15 | 132.44 | 121.56 | 122.63 | 130.13 |
| High | 124.31 | 115.56 | 112.63 | 122.75 | 107.06 | 106.00 |
| Trials | 135.50 | 115.38 | 122.53 | 122.16 | 114.84 | 118.06 |

[a]In this table, any contrast greater than 18.43 seconds is significant at or beyond .05 (Scheffe multiple comparison limits [Myers, 1966]).

TABLE 8

OBSERVED COMPARISONS OF VARIANCES ON

VERTICAL S-A MAXIMUM ALTITUDE

| | Experimental | | Treatment Control | Current | |
|---|---|---|---|---|---|
| | $A_1B_2$ | $A_1B_1$ | C | $A_2B_1$ | $A_2B_2$ |
| Variances | 5046.68 | 7032.68 | 28554.24 | 37419.03 | 52418.10 |
| Comparisons[a] | 1.00 | 1.37 | 5.66* | 7.41* | 10.39* |
| | | 1.00 | 4.06* | 5.32* | 7.45* |
| | | | 1.00 | 1.31 | 1.84 |
| | | | | 1.00 | 1.40 |

[a]1.00 as the leading value in a row indicates the variance in that column was used as the denominator in the *F*-ratio.

**F*-ratios significant at or beyond the predicted .05 values; each of these *F*-ratios would have been significant tested at the .001 level.

control group was not significantly less variable in performance at maximum altitude than the groups receiving current instruction.

Further analyses of variances for trials and subjects provide additional evidence that differences in variability may be important indicators of differences of quality among pilot performances. Cochran's test, $\underline{C}$, is commonly used to test for heterogeneity of variance prior to an ANOVA (Myers, 1966; Winer, 1971). As a test value, $\underline{C}$ is defined as the ratio of the largest variance to the sum of all the variances in a set. Conceptually, $\underline{C}$ represents the percent of variance accounted for by the largest variance in a set of variances.

To illustrate, for the variances tested in Table 8, the variance for group $A_2B_2$, 52418.10, accounts for 40% of the total variance. On the other hand, the combined variances (12079.36) for the groups receiving experimental instructions ($A_1B_1$ and $A_1B_2$) account for only 9% of the total. Among the possible tests of $\underline{C}$ for all trials and all subjects, single variances were identified that accounted for as much as 90% of the total variance in some groups. For example, among all possible comparisons of variances for each subject's performances, both the largest (361802.25) and the next to smallest (231.34) were located in one treatment group, $A_2B_1$. Among this set of all student pilot variances, the smallest was 187.42; compared to the variances of the two experienced pilots, 749.12 and 743.65, this smallest student pilot variance is not significant. A variance of 5228.86 or larger is required to be significant, .05, when compared to the experienced pilots' variances.

Discussion

Performance time and maximum altitude enable one to discriminate between performances of individuals and experimental treatments. To the extent that precision of pitch and power control in the Vertical S-A is reflected in variations of performance at maximum altitude, these findings provide strong support for the hypothesis that specific indicators should be considered. Brecke (1975) found that the groups receiving formal instructions, experimental or current, required about twice as much study time as the control group. He also failed to find significant differences between control group and treatment group performances on his dependent variables. On this evidence, he questioned the time and effort required to prepare experimental instruction. In the present investigation, performances of groups receiving the experimental instructions were about four to five times less variable than those of the control group and about seven times less variable than those of groups receiving current instruction. On the basis of this evidence, it is conceivable that a reduction of four or more times in variability of performance early in training is a highly desirable outcome.

## Design of the Main Investigation

In the second preliminary investigation, it was shown that analyses of specific indicators, i.e., variance at maximum altitude in the Vertical S-A, would lead to different results from analyses carried out on summary indicators like hit rate, percent time on criterion, and error amplitude. Two alternatives were available in the design of this

third investigation: (a) to identify particular locations of differences in the student pilot performances of Brecke's 1975 study, or (b) to examine relationships between summary and specific indicators empirically. To identify particular locations of differences, methods described in the performance state evaluation model would have been used. The first alternative was rejected because the utility of the performance state model had been demonstrated in the first preliminary investigation.

The main investigation was designed to examine relationships between summary and specific indicators. The hypothesis was that a set of selected specific indicators would predict values of a summary indicator. If summary indicator scores from Brecke (1975) data could be predicted from the selected specific indicators, the results of the main investigation would support the replacement of summary indicators with selected simple indicators. A multiple regression analysis was designed to test this hypothesis.

## Method

Design. Because error amplitude, E(T), is theoretically an indicator of variability in pilot performances, it was selected as the criterion variable in the multiple regression analysis. Brecke (1975) used ANOVA to analyze scores on error amplitude. The results of his analysis are given in Table 9 (Brecke, 1975, p. 64). In the present study, Brecke's E(T) scores were used as the criterion. The presence of significant effects in his ANOVA (Table 9) and the results of the second preliminary investigation in the present study led to a decision

TABLE 9

ANALYSIS OF VARIANCE: ERROR AMPLITUDE $E_{\Sigma}$[a]

| Source | SS | df | MS | F | p |
|---|---|---|---|---|---|
| Total | 2436.03 | 191 | 12.75 | | |
| Between | 1112.66 | 31 | 35.89 | | |
| Instruct. cues (A) | 115.20 | 1 | 115.20 | 3.38 | .0735 |
| Practice (B) | 21.41 | 1 | 21.41 | .63 | .5593 |
| A x B | 21.26 | 1 | 21.26 | .62 | .5577 |
| Error | 954.79 | 28 | 34.10 | | |
| Within | 1323.37 | 160 | 8.27 | | |
| Trials (D) | 119.52 | 5 | 23.90 | 3.14 | .0104 |
| D x A | 86.30 | 5 | 17.26 | 2.26 | .0508 |
| D x B | 13.09 | 5 | 2.62 | .34 | .8856 |
| D x A x B | 37.07 | 5 | 7.41 | .97 | .5619 |
| Error | 1067.39 | 140 | 7.62 | | |

[a]"$E_{\Sigma}$" signifies that error amplitude scores, E(T), were summed across five performance variables: airspeed, heading, vertical rate, pitch, and power (Brecke, 1975, p. 6).

to carry out a regression analysis on each performance trial with correlations averaged across treatment groups.

Performance state times, total time, maximum altitude, and measures of prior experience were used as the predictors or variates. Three combinations of performance state times were also included as predictors in the design. These combinations were the sums of performance state times 1 to 3, 5 tc 7, and 1 to 7. The prior experience variates were total hours as a pilot, flying time in training, total hours in the T-4 simulator, and number of minutes to complete an instructional program prior to performance of the Vertical S-A in a T-4G flight simulator. The regression design is summarized and the labels used for the variables are given in Table 10.

Data. Data used in this investigation was obtained from a previous study by Brecke (1975). In that study, 39 student pilots in UPT studied one of five sets of materials prior to performing a sequence of six trials on the Vertical S-A. In addition to data collected on six simulator performance variables and time, Brecke obtained data on the time taken by each subject to study the training materials and data on the prior experience of each student pilot. He obtained E(T) scores from the data collected with ADCS during each performance using methods developed by Shipley, et al. (1974). The E(T) scores used in Brecke's ANOVA were summed across five of the six simulator performance variables: airspeed, heading, vertical rate, pitch, and power.[2]

---

[2]The Brecke analysis did not include altitude data because ADCS malfunctions during the collection process made it impossible to use a computer to compute scores from the altitude data.

TABLE 10

SUMMARY OF VARIABLES IN DESIGN OF REGRESSION ANALYSIS

| Variable Number | Variable Type | Variable Name | Mnemonic |
|---|---|---|---|
| 1 | Criterion | Error Amplitude, Summed | Error |
| 2 | Performance | Total Time, transformed | TOTTIM |
| 3 | Performance | State One Time, transformed | Time 1 |
| 4 | Performance | State Two Time, transformed | Time 2 |
| 5 | Performance | State Three Time, transformed | Time 3 |
| 6 | Performance | State Four Time, transformed | Time 4 |
| 7 | Performance | State Five Time, transformed | Time 5 |
| 8 | Performance | State Six Time, transformed | Time 6 |
| 9 | Performance | State Seven Time, transformed | Time 7 |
| 10 | Performance | Maximum Altitude, transformed | MAXALT |
| 11 | Composite | Sum Times 1-3 | Sum 1 |
| 12 | Composite | Sum Times 5-7 | Sum 2 |
| 13 | Composite | Sum Times 1-7 | Sum 3 |
| 14 | Personal | Total Hours as Pilot | TOTHRS |
| 15 | Personal | Flying Time in UPT | UPTHRS |
| 16 | Personal | Time in T-4 Simulator | SIMHRS |
| 17 | Personal | Study Time (Minutes) | Study |

Procedures

Time and maximum altitude data, obtained in the second preliminary investigation, were transformed prior to the regression analysis. The transformations were made to account for possible offsetting compensations on performance times and to obtain normalized data. First, the standard values for time (Table 2) and 16,000 feet for maximum altitude were subtracted. Second, the absolute value of each difference was divided by the associated population standard deviation (Table 4). In the Brecke (1975) study, error amplitude scores were computed as standardized deviations from performance limits about the standard flight path.

After the data transformations, correlation matrices were obtained for each treatment group in the design of the Brecke study on each of the six performance trials. Each correlation in the set of 30 matrices was then transformed to a Fisher's $Z$ (Hays, 1963), the $Z$ values were averaged cross the five matrices for each trial, the means were re-converted to correlations, and the mean correlations were analyzed with a stepwise multiple regression. Trial effects, seen as an improvement in performance across trials, were present in the error amplitude scores. Twelve separate regression analyses, two on each trial, were used to avoid confounding correlations between criterion and predictors with the trial effects.

A standard stepwise multiple regression program in the Statistical Package for the Social Sciences (Nie, Bent, & Hull, 1970) was used for two different analyses of each matrix of mean correlations for each trial: (a) the variables designated as performance or composite in the design (Nos. 2-13), and (b) all variables (Table 10).

Conventional stepwise regression options set in the program were used: no limit on number of steps (maximum allowable is 80); a .01 level to add a given variable to the regression equation; and a .001 tolerance of linear relationship between a variable to be added and those already in the equation. Sample size was set at 34 in each analysis. At each step in the analysis, the output included an ANOVA summary, a summary of the relative contributions of each variable selected, a summary of the relative potential contribution of the remaining partial correlations, and a listing of the beta and b weights.

# CHAPTER IV

# RESULTS

This chapter is a report of tests of the hypothesis that a select set of specific indicators could be used to replace summary indicators. Twenty-four multiple regression analyses and one correlation between trial means were used to test this hypothesis. Support for the hypothesis would be obtained if two criteria were satisfied: (a) if a small set (4 to 6) of specific indicators were consistently selected in the regression equations; and (b) if equations using these indicators accounted for a substantial proportion (50% or more) of the variance in the criterion. In this chapter, the results of 12 planned regression analyses are reported and the results of 12 post hoc regression analyses are presented. Finally, results of a post hoc correlation between criterion means across trials and means of a selected specific indicator, maximum altitude, are reported.

## Results of Planned Analyses

Two sets of stepwise multiple regressions were carried out on mean correlations from each of six performance trials. Error amplitude, an objective indicator of variability in pilot performance data, was used as the criterion in all analyses. One analysis consisted of 12 variables taken from performance data. In the second analysis, four variables from student pilots' personal data were added to the set of

variables available for selection. In the present section, the results of the 12 regression analyses are summarized.

Regression on Performance Indicators

Of the 16 variables in the complete design (Table 10), 12 were indicators taken from performance data (9 predictors) or sums[3] of these indicators (3 predictors). These 12 variables were used in the first analyses because of their relationship to a performance state evaluation model of the standard Vertical S-A flight path (Figure 2 and Table 2). To the extent that performance states are a valid model and variables such as performance time or deviations from maximum altitude are valid as indicators of performance quality, a subset of these 12 indicators should account for a substantial proportion of the variance in the criterion, error amplitude. To test this hypothesis, six stepwise multiple regressions (SPSS program) were carried out with each variable given equal weight.

In Table 11, a summary is given of the variables selected in each analysis, their order of selection, and proportion of variance accounted for at two locations in each equation: (a) after the first five variables, and (b) at the end of the complete equation. Five variables were the least that would predict 50% of the variance in every equation. With no more than five variables, the equations accounted for at least 53% of the variance; the range was from 53% to 96%. With all the variables selected, the proportion of variance

---

[3] In this design, sums were justified because absolute deviations from standard values were used. These sums were included to test for two possible relationships: (a) compensating differences in total time, and (b) departure from symmetry.

TABLE 11

SUMMARY OF FIRST FIVE VARIABLES IN EACH TRIAL EQUATION SELECTED FROM TWELVE PERFORMANCE INDICATORS IN FIRST ANALYSIS

| Variable | | Trial | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Design Number | Mnemonic[a] | 1 | 2 | 3 | 4 | 5 | 6 | *f* |
| 2 | TOTTIM | | 3 | 1 | | | 1 | 3 |
| 3 | TIME 1 | | | | | | | |
| 4 | TIME 2 | 5 | 1 | | 1 | 1 | | 4 |
| 5 | TIME 3 | | 4 | 5 | | 2 | 2 | 4 |
| 6 | TIME 4 | | | | | | | |
| 7 | TIME 5 | 2 | | | | | 3 | 2 |
| 8 | TIME 6 | 4 | 2 | | | | | 2 |
| 9 | TIME 7 | | | | 3 | | | 1 |
| 10 | MAXALT | | | 3 | 2 | 3 | | 3 |
| 11 | SUM 1 | 3 | | 2 | 4 | 5 | 4 | 5 |
| 12 | SUM 2 | | 5 | | | 4 | 5 | 3 |
| 13 | SUM 3 | 1 | | 4 | 5 | | | 3 |
| Number in Complete Equation: | | 8 | 6 | 7 | 7 | 9 | 8 | |
| Proportion of Variance: | | | | | | | | |
| First 5 variables: | | .53 | .93 | .83 | .65 | .77 | .96 | |
| Complete equation: | | .55 | .93 | .90 | .82 | .81 | .97 | |

[a]Names of each variable are given in Table 10.

ranged from 55% to 97%. The increase beyond five variables was small or negligible except for the Trial 4 equation with a 17% increase. Among the 12 possible predictors, 2 were never selected among the first 5, while 7 were selected 3 or more times. To summarize, 58% (7 of 12) of the variables accounted for 83% (25 of 30) of the possible selections. Detailed results, i.e., ANOVAs and summary tables of the equations, are contained in Appendix A.

At this point, it was not possible to determine precisely how well the seven variables selected most frequently would perform as predictors. That is, the decision to use frequency of selection (in three or more equations) did not include information about order of selection or relative contribution of each variable selected to an equation. For example, one variable, Time 5, was selected in two regression equations as second or third predictor. This variable was associated with moderate (.11) to substantial (.32) increases in the proportion of variance accounted for. A five-step regression analysis was designed to test this subset of seven most frequently selected variables. The results of this post hoc analysis are discussed later in this chapter.

## Regression on All Variables

In addition to performance indicators, one would expect that measures of prior or related experience might be correlated with quality of performance early in training. Brecke (1975) examined this hypothesis as a possible basis for using analysis of covariance rather than ANOVA to analyze his dependent variables. He correlated measures of prior or related personal experience with scores on error amplitude,

hit rate, and percent time on criterion. Although some of the correlations were significant, .05, Brecke concluded that none were sufficiently large, i.e., $r \geq .60$, to include these measures in his design as covariates. Such measures might make moderate to substantial contributions to proportion of variance accounted for in a multiple regression analysis. In the present analysis, four measures from student pilots' personal data were added to the 12 performance indicators (Table 10) and a second set of stepwise regression equations were developed.

A summary[4] of the variables selected is given in Table 12. Criteria similar to those in the previous analysis were used: consideration of the first five variables selected in each equation and proportion of variance accounted for by each set of five variables. Each measure from personal data appeared in at least one regression equation. Total pilot hours (TOTHRS) and study time (STUDY) were selected two or more times. Relative to performance indicators alone (Table 11), measures from personal data improved equation effectiveness with one exception. In the equation for Trial 2, simulator hours (SIMHRS) was selected as the best predictor instead of Time 2 and the net effect was a 12% decrease in proportion of variance at the fifth step.

These are nearly optimum equations and these $R^2$ values would be expected to shrink when used with other data. As an alternative to cross validation, a restricted set of nine variables was selected and a second five-step analysis was carried out. Three of the 16 variables

[4]As in Table 11, order of selection is presented along with the number of variables and the proportion of variance due to regression for the complete equation.

TABLE 12

SUMMARY OF FIRST FIVE VARIABLES IN EACH TRIAL EQUATION SELECTED FROM SIXTEEN VARIABLES IN COMPLETE DESIGN

| Variable Design Number | Mnemonic[a] | Trial 1 | 2 | 3 | 4 | 5 | 6 | f |
|---|---|---|---|---|---|---|---|---|
| 2 | TOTTIM | | 4 | 1 | | | 1 | 3 |
| 3 | TIME 1 | | | | | | | |
| 4 | TIME 2 | | 2 | | 1 | 1 | | 3 |
| 5 | TIME 3 | 3 | 5 | 5 | 5 | 2 | 2 | 6 |
| 6 | TIME 4 | | 3 | | | | | 1 |
| 7 | TIME 5 | | | | | | 3 | 1 |
| 8 | TIME 6 | | | | | | | |
| 9 | TIME 7 | | | | | | | |
| 10 | MAXALT | | | 3 | | 3 | | 2 |
| 11 | SUM 1 | | | 2 | 4 | | 4 | 3 |
| 12 | SUM 2 | 4 | | | | 4 | | 2 |
| 13 | SUM 3 | 1 | | 4 | | | | 2 |
| 14 | TOTHRS | 2 | | | 3 | | 5 | 3 |
| 15 | UPTHRS | 5 | | | | | | 1 |
| 16 | SIMHRS | | 1 | | | | | 1 |
| 17 | STUDY | | | | 2 | 5 | | 2 |
| Number in Complete Equation: | | 7 | 9 | 9 | 9 | 11 | 7 | |
| Proportion of Variance: | | | | | | | | |
| First 5 variables: | | .97 | .81 | .82 | .94 | .77 | .98 | |
| Complete equation: | | .99 | .97 | .99 | 1.00 | .97 | 1.00 | |

[a]Names of each variable are given in Table 10.

were not selected among the first five in any of the six equations in the present analysis; two of these (Time 1 and Time 7) were not selected in the previous analysis. Four of the remaining 13 variables were selected only once. These seven were eliminated leaving a subset of 9 to be used in a post hoc analysis described later.

Time 3 was selected in ten equations between the two analyses. It was never selected higher than second; but as a lower order predictor it accounted for rather large increases in proportions of variance, range .10 to .36. The simple correlations between Time 3 and the criterion ranged from -.36 to .49. In combination with other variables, Time 3 appears as an important predictor variable. Detailed results, i.e., ANOVAs and summary tables of the regression equations, are given in Appendix B.

To summarize, it appeared that three variables (Time 2, Time 3, and Sum 1) might well be used as the basis of a five variable equation for all trials. This possibility was not tested directly (by forcing these three variables into each equation prior to any others) because of the pattern of selections of best and second best predictors and of patterns among the first order correlations. Further, problems of heterogeneity of variances were known to be present in the data (as a result of the second preliminary investigation) and, consequently, an effort to obtain a "best" equation seemed unwarranted. Instead, the subset of nine most frequently selected variables was submitted to a five-step regression analysis.

## Results of Five-Step Regressions

From the two preceding regression analyses, it was possible to develop restricted subsets of the variables. Subsets of seven and nine variables were developed by eliminating infrequently used variables. Elimination was based on frequency of use and did not include information about relative contributions to proportion of variance. To determine how well equations of not more than five variables from these two subsets would account for variance in the criterion, two additional sets of multiple regressions were carried out. Each regression was limited to five steps and was carried out on each performance trial. The results were 12 five variable regression equations.

### Regression on Seven Performance Variables

A subset of seven performance variables formed the basis of the first six equations. A summary of the results of these six equations, one for each performance trial, is given in Table 13. With the exception of the equation for Trial 1, each equation accounted for at least 50% of the variance in the criterion (range was 34% to 82%). Among the variables, maximum altitude (MAXALT) was included in each equation in the present analysis.

In the first analysis (on performance variables), MAXALT was included in only three equations (Table 11). MAXALT showed the greatest change in number of inclusions but it had a low mean order of inclusion (3.67). Sum 2 was included twice in the present analysis, as compared to three inclusions previously, and it had the lowest mean order (4.5). On this basis, Sum 2 appears to be a marginal indicator, whereas the

TABLE 13

SUMMARY OF FIVE-STEP REGRESSION: SUBSET OF
SEVEN PERFORMANCE VARIABLES

| Variable | | Equations on Trials | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Design Number | Mnemonic[a] | 1 | 2 | 3 | 4 | 5 | 6 | f |
| 2 | TOTTIM | 3 | | 1 | 4 | | 1 | 4 |
| 4 | TIME 2 | 4 | 1 | | 1 | 1 | | 4 |
| 5 | TIME 3 | | 3 | 5 | | 2 | 2 | 4 |
| 9 | MAXALT | 5 | 5 | 3 | 2 | 3 | 4 | 6 |
| 10 | SUM 1 | 2 | 2 | 2 | 5 | 5 | | 5 |
| 11 | SUM 2 | | | | | 4 | 5 | 2 |
| 12 | SUM 3 | 1 | 4 | 4 | 3 | | 3 | 5 |
| Proportion of Variance: | | .34 | .55 | .82 | .58 | .77 | .63 | |

[a]Names for each variable are given in Table 10.

effects of MAXALT are much less clear. More consideration is given to MAXALT after a consideration of the second set of five-step regression equations. Detailed tables of ANOVAs and the equations obtained are given in Appendix C.

Regression on Nine Variables

A subset of nine variables was used in the second five-step regression analysis. In addition to the subset of seven performance variables, two variables (TOTHRS and STUDY) were added from the set of personal experience variables. The data in Table 14 show substantial increases in the proportions of variances compared to those in Table 13; these increases were: Trial 1, .53; Trial 2, .33; and Trial 4, .36. There were no differences in proportions on Trial 3 (same equation for both analyses) and Trials 5 and 6 (with one variable different at the fifth step in each). Detailed results of ANOVAs and summaries for each of the five-step multiple regression equations are given in Appendix D.

Summary

On frequency of inclusion, a shift was observed between Time 3 and MAXALT. As seen in Tables 13 and 14, Time 3 increases from four to six inclusions, while MAXALT decreases from six to four. Time 2 and TOTTIM are consistently selected as first variable. At first, the implications of this shifting were not entirely clear. That is, in all four analyses, and especially those summarized in Tables 12 and 14, equations differed across trials by kind of variable included. Early in the performance, i. e., Trials 1 and 2, equations consisted of

TABLE 14

SUMMARY OF FIVE-STEP REGRESSION:
SUBSET OF NINE VARIABLES

| Variable | | Equations on Trials | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Design Number | Mnemonic[a] | 1 | 2 | 3 | 4 | 5 | 6 | f |
| 2 | TOTTIM | | | 1 | | | 1 | 2 |
| 4 | TIME 2 | | 1 | | 1 | 1 | | 3 |
| 5 | TIME 3 | 3 | 3 | 5 | 5 | 2 | 2 | 6 |
| 9 | MAXALT | 5 | | 2 | | 2 | 4 | 4 |
| 10 | SUM 1 | | 2 | 2 | 4 | | | 3 |
| 11 | SUM 2 | 4 | | | | 4 | | 2 |
| 12 | SUM 3 | 1 | 4 | 4 | | | 5 | 4 |
| 13 | TOTHRS | 2 | | | 3 | | 3 | 3 |
| 16 | STUDY | | 5 | | 2 | 5 | | 3 |
| Proportion of Variance: | | .87 | .88 | .82 | .94 | .77 | .62 | |

[a]Names of each variable are given in Table 10.

composite performance variables (Sums) and prior experience variables. Later in performance, Trials 5 and 6, specific performance indicators, i.e., Time 2, Time 3, and TOTTIM, are included more frequently overall and more frequently as first to third. Review of the outcomes of these equations (Appendices A to D) revealed that on Trials 5 and 6, the first 2 or 3 variables would account for at least 50% of the variance (Trial 5: 51% with 2, 65% with 3 variables; Trial 6: 55% with 2, 57% with 3 variables).

Two facts about the performance data would possibly account for the shift of variables in equations across trials. First, there were problems of heterogeneity of variance on at least some variables as well as problems of measurement error and error due to ADCS malfunctions. It was known that the ADCS malfunctions were not uniformly distributed across subjects or trials. Second, and more important, the data contained evidence of experimental treatment effects and effects of improvement due to learning across trials. In particular, it was known that error amplitude means were curvilinear across trials and this fact led to a final analysis.

## Analysis of Means Across Trials

The trial means of the criterion variable, error amplitude, exhibited a definite curvilinear pattern. An array of means was designed to examine possible relationships between trial means of the criterion and each of the performance variables (Table 15). Simple visual inspection revealed that the means of MAXALT (standard deviates, i.e., *z*-scores, from the standard of 16,000 feet) were also curvilinear

TABLE 15

MEANS OF ERROR AMPLITUDE AND TWELVE SPECIFIC
PERFORMANCE INDICATORS ACROSS TRIALS

| | Trial | | | | | |
|---|---|---|---|---|---|---|
| Variable | One | Two | Three | Four | Five | Six |
| *Error | 7.24 | 5.84 | 5.42 | 4.33 | 5.60 | 5.90 |
| TOTTIM | 2.12 | 2.22 | 1.78 | 1.85 | 1.88 | 1.77 |
| TIME 1 | 2.91 | 2.68 | 2.55 | 2.47 | 2.34 | 2.26 |
| TIME 2 | 1.19 | 1.24 | 1.43 | 1.47 | 1.41 | 1.14 |
| TIME 3 | 3.28 | 2.50 | 3.02 | 2.72 | 1.25 | 2.31 |
| TIME 4 | 1.19 | 1.25 | 1.13 | .95 | 1.00 | .87 |
| TIME 5 | 1.39 | 1.61 | .82 | 1.03 | 1.09 | .99 |
| TIME 6 | 2.25 | 2.04 | 1.68 | 1.92 | 2.04 | 1.92 |
| TIME 7 | 5.90 | 3.60 | 2.25 | 2.85 | 1.94 | 2.08 |
| *MAXALT | 5.00 | 3.59 | 2.72 | 3.89 | 2.65 | 3.07 |
| SUM 1 | 7.38 | 6.42 | 6.99 | 6.66 | 5.00 | 5.71 |
| SUM 2 | 9.53 | 7.25 | 4.76 | 5.80 | 5.07 | 4.99 |
| SUM 3 | 18.10 | 14.92 | 12.89 | 13.41 | 11.07 | 11.58 |

in a pattern similar to those of error amplitude with one exception, Trial 4. The two sets of means were plotted (Figure 8) and the significance of the relationship was immediately obvious.

In the MAXALT means, the mean for Trial 4 deviated from the best fitting curve (visual fit). A test of the standard error of the mean on that trial revealed that a 95% confidence interval would include the best fitting curve. A new mean, 2.60, was interpolated for MAXALT on Trial 4. Two Pearson product-moment correlations were carried out on the two sets of means across trials. One correlation, $r = .98$, was computed with the interpolated mean for MAXALT and the other, $r = .72$, with the observed mean.

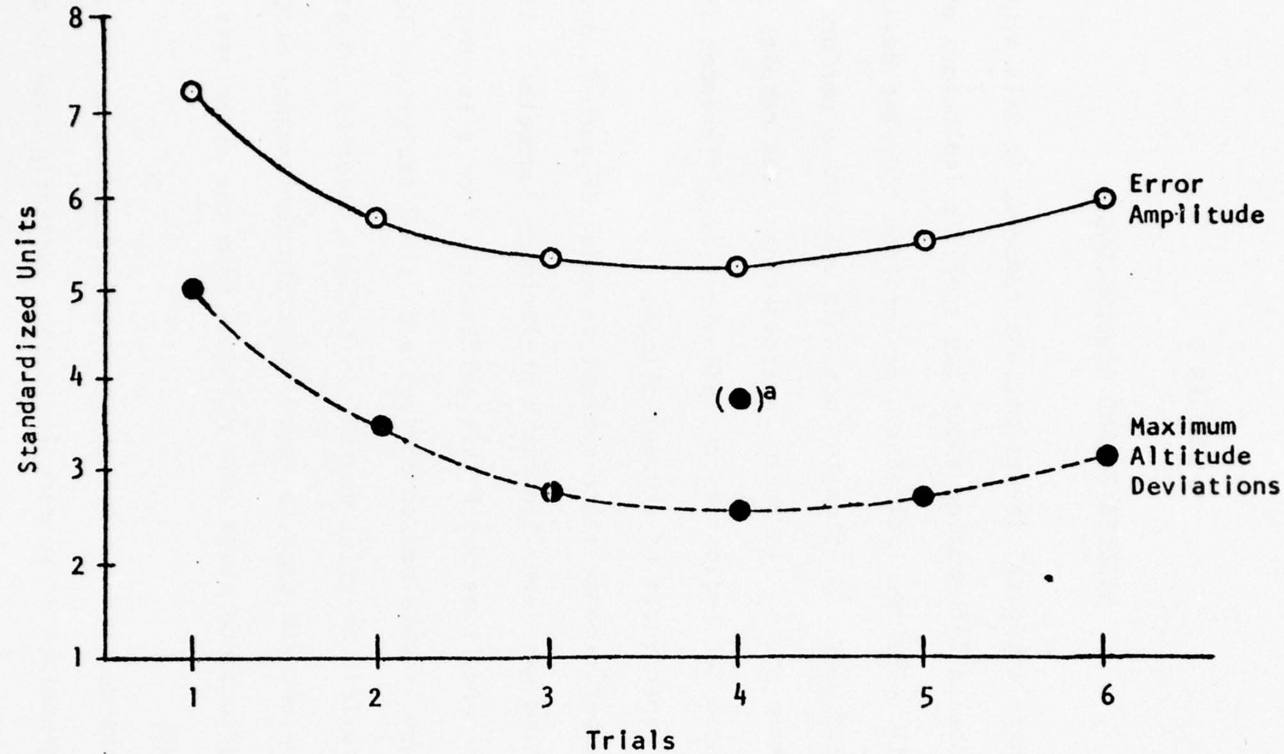


Figure 8.--Means of Error Amplitude and Maximum Altitude Deviations in Standardized Units Across Trials

---

[a]This outlier is the observed mean for maximum altitude deviations on Trial 4; 95% confidence limits included the interpolated value of 2.60 as shown.

# CHAPTER V

## DISCUSSION AND CONCLUSIONS

The three empirical investigations reported in this study were designed to obtain information about two specific indicators of performance skill. The two indicators, performance time and deviations from a standard, were selected for possible use with a performance state evaluation model. In these investigations, the results support the use of these two indicators as indicators of differences in performances of experienced or student pilots.

In the performance state evaluation model (Figure 7, page 36), performance time was identified as a preliminary indicator. If an overall performance time deviates significantly from a standard time value, evidence is obtained to conduct a detailed analysis. To conduct this detailed analysis, deviations between observed and assigned performance values are used to locate specific performance errors whenever a performance state time deviates from the associated standard time value.

Performance time and deviations from a standard were investigated as alternatives to summary indicators currently used in pilot training research and development. A summary indicator, e.g., error amplitude, was computed as a function of a sum from all observations in a set of time series data. Some objections to summary indicators of this nature were (a) lack of sensitivity to a few large deviations

as the number of observations increased, and (b) difficulty of application, i.e., the need for a high frequency of observations and computational complexity. The results of these investigations support the replacement of summary indicators with specific indicators.

There were two major findings. These were: (a) that, with data from an experiment, the specific indicators were more sensitive to the effects of differences in experimental treatments than were summary indicators; and (b) that a small set of specific indicators would account for moderate (34%) to a large (82%) proportions of the variance in error amplitude in a regression analysis. As an additional outcome of the regression analysis, it was found that trial means of a specific indicator, deviation from maximum altitude, exhibited the same curvilinear trend of improvement in performance as trial means for error amplitude.

These findings are interesting because they suggest that even with fewer data points on initial observation, the outcomes of evaluation will be superior to those obtained using summary indicators. In terms of the performance state evaluation model, it was found (a) that performance times would discriminate between individual performances of experienced pilots or between group performances in a training experiment; (b) that maximum altitude variances would discriminate between performances of groups in a training experiment; and (c) that two variables, Time 3 and maximum altitude, were consistently identified in the regression equations.

This last finding is especially interesting in terms of the performance state model. Time 3 represents the time from performance

State 3, transition from climb to maximum altitude in the Vertical S-A (Figure 2, page 12). Referring to the algorithm (Figure 7, page 36), the outcomes in these investigations would support the recommended analysis procedures. First, it was found that total times were different. Second, maximum altitude is the end point of performance State 3 and Time 3 was found to be a frequently selected variable in the regression equations.

Further research is needed to determine specifically how often this combination of outcomes will be observed in individual performances of the Vertical S-A. Research is also needed to determine other specific deviations that might be located in Vertical S-A data using performance time. The outcomes of the first investigation show that time will locate deviant performance states. The obvious relationship between mean performance trends on error amplitude and deviations at maximum altitude suggest the hypothesis that size of deviations on other performance variables, e.g., airspeed, will be correlated with performance states.

Within a performance state evaluation model, performance time and deviations from standard flight path values may also be used by instructor pilots. For example, maximum altitude in the Vertical S-A would be as easily observed by a human as by an automated system. The findings in these investigations would suggest that, for the operational user, maximum altitude in the Vertical S-A could replace an overall rating. Confirmation of this hypothesis would result in a single objective observation replacing a global subjective rating. To the extent that similar values can be identified in other training

AD-A033 920 ARIZONA STATE UNIV TEMPE COLL OF EDUCATION F/G 5/9
AN AUTOMATED MEASUREMENT TECHNIQUE FOR EVALUATING PILOT SKILL.(U)
FEB 76 B D SHIPLEY AF-AFOSR-2900-76
UNCLASSIFIED TN-60229 AFOSR-TR-76-1253 NL

2 OF 2
ADA033920

END
DATE
FILMED
2 -77

maneuvers, a workable solution will be achieved for the dilemma of excessive detail versus uninformative generality in pilot training measurement and evaluation.

To summarize the outcomes of the present research, consider again the five objectives set out for a training research approach to measurement and evaluation studies in pilot training:

1. To identify potentially critical points or events in descriptions of pilot behaviors that make up the operational sequence and of the performance task.

2. To develop observation schedules and scoring procedures to account for the effects of these events on performance skill.

3. To determine empirically relative frequencies of these critical events throughout performances of the assigned task from objective data.

4. To train instructor pilots, check pilots, and other pilot training personnel to employ the schedules and procedures with student pilot performances, first in a simulator, then in the aircraft.

5. To develop reliability assessment procedures for use with measurement and evaluation practices in the aircraft based on the outcomes of the four preceding objectives.

In the present study, the first three objectives were investigated. In terms of the first objective, the outcomes from this study were: (a) that existing methods from a maneuver analysis, e.g., Brecke and Gerlach (1972), were a suitable basis to determine criteria and values for the purposes of developing a standard flight path; (b) that empirical methods must be used to establish estimates of the

standard flight path values; and (c) that critical points within an operational sequence are most likely to be located in the vicinities of transition states. Although the analytic methods have only been applied in the case of one instrument training maneuver, other researchers have effectively employed similar methods (Knoop & Welde, 1973). Nevertheless, research is needed to generalize these three findings to other pilot training maneuvers.

The second objective was the source of a dilemma which motivated much of the present research: how to obtain indicators of skill that were intermediate between excessive detail, e.g., high rate time sampling data from ADCS, and uninformative generality, e.g., global ratings or summary indicators. To the extent that one can generalize from the results in the present research, performance time and deviations from a standard can be effectively combined with a performance state evaluation model to solve this problem. In particular, it would appear that superior evaluations can be obtained with fewer data points on initial observation. In this area, more research is needed with other maneuvers to refine the procedures for preparing observation schedules from the performance state model and training objectives.

In the present study, the analysis algorithm (Figure 7) served as the basis to determine relative frequencies of critical events, i.e., objective three above. Results from the present investigations should be considered as evidence to support the use of this set of analytic procedures. The evidence for using these procedures is strongest at the first steps of the algorithm. Extensive diagnostic

analyses were not made because the characteristics of the available data were not considered adequate for such an effort and because the collection of new data was beyond the scope of the present study. Subsequent research in this area might well begin by collecting new data from both instructor and student pilot performances to carry out such detailed analyses. As an hypothesis, Shipley, Gerlach, and Brecke (1974) have suggested that differences in patterns of errors or forms of a performance in time might be used to develop a scheme to classify performances.

The last two objectives, four and five, were not included directly in the scope of the present research. Indirectly, a secondary objective was to develop measures and methods that could be used in the operational and management areas as well as in training research and development. By implication from parallel research on student pilot training methods, of which this measurement and evaluation study was a part, specific indicators can be combined with the algorithmic procedures and training criteria to develop an instructor pilot evaluation training program. The fifth objective cannot be effectively considered until the suggested research and development is completed and a training program is at least in a prototype form.

Another potential contribution of the present study was a practical application of tests for differences in variances in the area of training research. In the second investigation, tests for differences in variances were used to test for differences in the effects of experimental treatments. These differences in experimental effects were predicted from a theoretical analysis of performance requirements and the

outcomes of a previous study. Similar uses of variance tests should be investigated. Distributions of scores may differ in variability as well as in central tendency or even when there are no differences in central tendency. Although tests for differences in variances are legitimate (and possibly even highly informative), it is generally permissible to ignore them because the more commonly used tests for differences in means, e.g., *t*-tests and ANOVA, are robust under the conditions of moderate violations of homogeneity of variance (Myers, 1966; Winer, 1971). However, in cases where experimental treatments can be expected to influence variability of performance, as in pilot training, Winer recommends that variance tests be used.

It is conceivable that similar effects might be found in other areas of training and instruction. For example, differences in the effects of instructional programs might be better reflected as differences in variability of achievement than as differences in mean achievement. This potential effect would possibly be usable in cases of criterion referenced tests and measures of mastery of performance on the same task over time. In general, the more complex the task, the more likely that changes in variances will indicate changes in performances due to training.

To conclude, in previous research on measures of skill in human performance, Fitts, Bahrick, Briggs, and Noble (1959) made this observation:

> Of course, every study uses some response measures, but usually the main purpose of the study is to find out more about procedural, organismic, or task variables, and the response measure which reveals these effects is

> often chosen on the basis of convenience. The underlying assumptions in such instances are that response measures are well understood, and that the various possible indicants for a given process measure very nearly the same things, so that one can choose arbitrarily among them on the basis of convenience. (p. 6.1)

Later in the same study, these authors conclude "that our understanding of skilled performance depends upon the development of analytical indicants of performance" (p. 6.47). Surely our understanding of skilled <u>pilot</u> performance depends upon the development of analytical indicants of performance.

# REFERENCES

## REFERENCES


Brecke, F. H. Transfer from selected pretraining variables to a complex perceptual motor task. Unpublished doctoral dissertation, Arizona State University, 1975.

Brecke, F. H., & Gerlach, V. S. Model and procedures for an objective maneuver analysis (Tech. Rep. No. 21201). Tempe: Arizona State University, Instructional Resources Laboratory, December, 1972.

Brecke, F. H., Gerlach, V. S., & Shipley, B. D. Effects of instructional cues on complex skill learning (Tech. Rep. No. 40829). Tempe: Arizona State University, College of Education, August, 1974.

Christensen, J. M., & Mills, R. G. What does the operator do in complex systems? Human Factors, 1967, 9, 329-340.

Connelly, E. M., Schuler, A. R., & Knoop, P. A. Study of adaptive mathematical model for deriving automated pilot performance measurement techniques. *Volume 1: Model development* (Tech. Rep. TR-69-7, Volume 1). Air Force Systems Command, Wright Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, October, 1969.

Dickman, J. L. Automated performance measuring. Paper presented at the NTECI Industry Conference, Orlando, Florida, November, 1974.

Etkin, B. Dynamics of atmospheric flight. New York: John Wiley, 1972.

Fitts, P. M., Bahrick, H. P., Briggs, G. E., & Noble, M. E. Skilled performance (Final Report R. F. Project 686). Columbus: Ohio State University Research Foundation, 1959.

Gerlach, V. S., Brecke, F. H., Reiser, R., & Shipley, B. D. The generation of cues based on a maneuver analysis (Tech. Rep. No. 21202). Tempe: Arizona State University, Instructional Resources Laboratory, December, 1972.

Hagin, W. V. Objective measures of single-engine instrument flying skill at the basic level of flying training. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D. C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Hill, J. H., & Goebel, R. A. Development of automated GAT-1 performance measures (Tech. Rep. No. 71-18). Air Force Systems Command, Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, May, 1971.

Horner, W. R., Radinsky, R. L., & Fitzpatrick, R. The development, test, and evaluation of three pilot performance reference scales (Tech. Rep. No. 70-22). Air Force Systems Command, Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, August, 1970.

Jacobs, Q. L. R. Introduction to random processes. In D. Bell & A. W. J. Griffin (Eds.), Modern control theory and computing. New York: McGraw-Hill, 1969.

Knoop, P. A., & Welde, W. L. Automated pilot performance assessment in the T-37: A feasibility study (Tech. Rep. No. 76-6). Air Force Systems Command, Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, April, 1973.

Koonce, J. M. Effects of ground based simulator motion conditions upon prediction of aircraft pilot proficiency. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1974.

Leshowitz, B., & Nelsen, G. M. Adaptive procedure for measurement of flying training proficiency. Interim Scientific Report. Unpublished manuscript, Arizona State University, 1974-75.

Merrill, M. D. Fundamentals of instructional design. In M. D. Merrill (Ed.), Instructional design: Readings. Englewood Cliffs, N.J.: Prentice Hall, 1971.

Miller, N. E. The problem of measuring flying proficiency. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D.C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

Myers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

Nie, N., Best, D. H., & Hull, C. H. SPSS statistical package for the social sciences. New York: McGraw-Hill, 1970.

Reid, L. D., & Etkin, B. Human pilots and handling qualities. In B. Etkin, Dynamics of atmospheric flight. New York: John Wiley, 1972.

Reiser, R., Brecke, F. H., & Gerlach, V. S. On the difference between procedure and technique in pilot instruction (Tech. Rep. No. 21128). Tempe: Arizona State University, Instructional Resources Laboratory, November, 1972.

Shipley, B. D., Gerlach, V. S., & Brecke, F. H. Measurement of flight performance in a flight simulator (Tech. Rep. No. 40830). Tempe: Arizona State University, College of Education, August, 1974.

Thorndike, R. C. Research problems and techniques. Washington, D.C.: U. S. Government Printing Office, Army Air Forces Aviation Psychology Program, Research Report No. 3, 1947.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Youtz, R. P., & Ericksen, S. C. Analysis of the pilot's task. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D.C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

# APPENDICES

# APPENDIX A

TABLE A1

STEPWISE REGRESSION FOR TRIAL ONE: PERFORMANCE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | SUM 3 | Regression | 1 | 87.98 | 13.91 | .303 | .202 | .942 |
| | | Residual | 32 | 6.47 | | | | |
| 2 | TIME 5 | Regression | 2 | 60.84 | 10.76 | .410 | .882 | .588 |
| | | Residual | 31 | 5.66 | | | | |
| 3 | SUM 1 | Regression | 3 | 44.34 | 8.11 | .448 | -.364 | -.727 |
| | | Residual | 30 | 5.47 | | | | |
| 4 | TIME 6 | Regression | 4 | 35.87 | 6.78 | .483 | -.739 | -.492 |
| | | Residual | 29 | 5.29 | | | | |
| 5 | TIME 2 | Regression | 5 | 31.24 | 6.21 | .526 | .690 | .230 |
| | | Residual | 28 | 5.03 | | | | |
| 6 | TOTTIM | Regression | 6 | 27.01 | 5.40 | .546 | .281 | .188 |
| | | Residual | 27 | 5.00 | | | | |
| 7 | MAXALT | Regression | 7 | 23.44 | 4.59 | .553 | .039 | .103 |
| | | Residual | 26 | 5.11 | | | | |
| 8 | TIME 1 | Regression | 8 | 20.54 | 3.87 | .553 | -.138 | -.046 |
| | | Residual | 25 | 5.31 | | | | |
| | Constant | | | | | | 5.472 | |

TABLE A2

STEPWISE REGRESSION FOR TRIAL TWO: PERFORMANCE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TIME 2 | Regression | 1 | 52.70 | 6.90 | .177 | 6.985 | 2.328 |
| | | Residual | 32 | 7.63 | | | | |
| 2 | TIME 6 | Regression | 2 | 40.00 | 5.71 | .269 | 8.685 | 2.895 |
| | | Residual | 31 | 7.00 | | | | |
| 3 | TOTTIM | Regression | 3 | 42.46 | 7.51 | .429 | -4.578 | -3.052 |
| | | Residual | 30 | 5.65 | | | | |
| 4 | TIME 3 | Regression | 4 | 57.19 | 24.30 | .770 | 1.035 | 1.725 |
| | | Residual | 29 | 2.35 | | | | |
| 5 | SUM 2 | Regression | 5 | 55.03 | 70.53 | .926 | - .546 | - .728 |
| | | Residual | 28 | .78 | | | | |
| 6 | SUM 1 | Regression | 6 | 46.05 | 60.14 | .930 | .162 | .325 |
| | | Residual | 27 | .77 | | | | |
| | Constant | | | | | | -9.455 | |

TABLE A3

STEPWISE REGRESSION FOR TRIAL THREE: PERFORMANCE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTTIM | Regression | 1 | 17.22 | 4.80 | .130 | .307 | .153 |
| | | Residual | 32 | 3.59 | | | | |
| 2 | SUM 1 | Regression | 2 | 13.29 | 3.91 | .201 | -.945 | -3.781 |
| | | Residual | 31 | 3.40 | | | | |
| 3 | MAXALT | Regression | 3 | 13.10 | 4.24 | .298 | 1.209 | 1.813 |
| | | Residual | 30 | 3.09 | | | | |
| 4 | SUM 3 | Regression | 4 | 18.90 | 9.72 | .573 | .768 | 3.840 |
| | | Residual | 29 | 1.94 | | | | |
| 5 | TIME 3 | Regression | 5 | 21.75 | 26.22 | .824 | -.353 | -1.237 |
| | | Residual | 28 | .83 | | | | |
| 6 | TIME 6 | Regression | 6 | 19.65 | 37.71 | .893 | -.532 | - .266 |
| | | Residual | 27 | .52 | | | | |
| 7 | TIME 5 | Regression | 7 | 16.98 | 33.66 | .901 | -.283 | - .141 |
| | | Residual | 26 | .50 | | | | |
| | Constant | | | | | | -.199 | |

TABLE A4

STEPWISE REGRESSION FOR TRIAL FOUR: PERFORMANCE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TIME 2 | Regression | 1 | 61.26 | 27.71 | .464 | 1.660 | .830 |
| | | Residual | 32 | 2.21 | | | | |
| 2 | MAXALT | Regression | 2 | 35.29 | 17.80 | .535 | .276 | .965 |
| | | Residual | 31 | 1.98 | | | | |
| 3 | TIME 7 | Regression | 3 | 25.46 | 13.74 | .579 | -1.706 | -1.706 |
| | | Residual | 30 | 1.85 | | | | |
| 4 | SUM 1 | Regression | 4 | 19.46 | 10.43 | .590 | -1.299 | -2.598 |
| | | Residual | 29 | 1.87 | | | | |
| 5 | SUM 3 | Regression | 5 | 17.18 | 10.43 | .651 | 1.051 | 3.152 |
| | | Residual | 28 | 1.65 | | | | |
| 6 | TIME 6 | Regression | 6 | 17.90 | 19.64 | .814 | -1.577 | - .788 |
| | | Residual | 27 | .91 | | | | |
| 7 | TOTTIM | Regression | 7 | 15.51 | 17.20 | .822 | .263 | .132 |
| | | Residual | 26 | .90 | | | | |
| | Constant | | | | | | 5.413 | |

TABLE A5

STEPWISE REGRESSION FOR TRIAL FIVE: PERFORMANCE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | $F$ | $R^2$ | $b$ | Beta |
| 1 | TIME 2 | Regression | 1 | 99.23 | 16.06 | .334 | 3.508 | 1.169 |
| | | Residual | 32 | 6.18 | | | | |
| 2 | TIME 3 | Regression | 2 | 75.64 | 16.09 | .509 | -1.184 | - .395 |
| | | Residual | 31 | 4.70 | | | | |
| 3 | MAXALT | Regression | 3 | 64.80 | 18.94 | .655 | .925 | .617 |
| | | Residual | 30 | 3.42 | | | | |
| 4 | SUM 2 | Regression | 4 | 54.68 | 20.26 | .736 | 1.859 | 1.240 |
| | | Residual | 29 | 2.70 | | | | |
| 5 | SUM 1 | Regression | 5 | 45.65 | 18.59 | .768 | -1.427 | - .951 |
| | | Residual | 28 | 2.46 | | | | |
| 6 | TIME 6 | Regression | 6 | 38.83 | 16.37 | .784 | -1.824 | - .608 |
| | | Residual | 27 | 2.37 | | | | |
| 7 | TOTTIM | Regression | 7 | 34.04 | 15.08 | .802 | .360 | .120 |
| | | Residual | 26 | 2.26 | | | | |
| 8 | TIME 5 | Regression | 8 | 29.98 | 13.11 | .807 | -1.269 | - .423 |
| | | Residual | 25 | 2.29 | | | | |
| 9 | TIME 4 | Regression | 9 | 26.87 | 11.68 | .814 | - .361 | - .120 |
| | | Residual | 24 | 2.30 | | | | |
| | Constant | | | | | | 3.299 | |

TABLE A6

STEPWISE REGRESSION FOR TRIAL SIX: PERFORMANCE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTTIM | Regression | 1 | 248.12 | 13.76 | .301 | 4.092 | .818 |
| | | Residual | 32 | 18.03 | | | | |
| 2 | TIME 3 | Regression | 2 | 225.47 | 18.69 | .547 | 2.460 | 1.968 |
| | | Residual | 31 | 12.07 | | | | |
| 3 | TIME 5 | Regression | 3 | 237.48 | 63.29 | .864 | -5.787 | -1.157 |
| | | Residual | 30 | 3.75 | | | | |
| 4 | SUM 1 | Regression | 4 | 192.92 | 104.92 | .936 | - .816 | - .653 |
| | | Residual | 29 | 1.84 | | | | |
| 5 | SUM 2 | Regression | 5 | 157.72 | 121.36 | .956 | - .849 | - .340 |
| | | Residual | 28 | 1.30 | | | | |
| 6 | TIME 6 | Regression | 6 | 133.04 | 134.20 | .968 | 1.133 | .227 |
| | | Residual | 27 | .99 | | | | |
| 7 | TIME 4 | Regression | 7 | 114.76 | 137.79 | .974 | .608 | .122 |
| | | Residual | 26 | .83 | | | | |
| 8 | MAXALT | Regression | 8 | 100.50 | 119.56 | .975 | - .074 | - .045 |
| | | Residual | 25 | .84 | | | | |
| | Constant | | | | | | 5.171 | |

APPENDIX B

TABLE B1

STEPWISE REGRESSION FOR TRIAL ONE: ALL VARIABLES

| Step | Variable | Analysis of Variance: Source | df | MS | $F$ | Summary Values: $R^2$ | $b$ | Beta |
|---|---|---|---|---|---|---|---|---|
| 1 | SUM 3 | Regression | 1 | 89.98 | 13.91 | .303 | 1.795 | 8.378 |
| | | Residual | 32 | 6.47 | | | | |
| 2 | TOTHRS | Regression | 2 | 70.14 | 13.87 | .472 | -1.069 | -1.069 |
| | | Residual | 31 | 5.06 | | | | |
| 3 | TIME 3 | Regression | 3 | 61.68 | 16.53 | .623 | -2.124 | -3.541 |
| | | Residual | 30 | 3.73 | | | | |
| 4 | SUM 2 | Regression | 4 | 62.07 | 36.94 | .836 | -1.558 | -5.193 |
| | | Residual | 29 | 1.68 | | | | |
| 5 | UPTHRS | Regression | 5 | 57.63 | 181.92 | .970 | - .423 | - .423 |
| | | Residual | 28 | .32 | | | | |
| 6 | TIME 5 | Regression | 6 | 43.63 | 251.27 | .982 | - .557 | - .371 |
| | | Residual | 27 | .19 | | | | |
| 7 | MAXALT | Regression | 7 | 42.02 | 384.57 | .990 | .047 | .124 |
| | | Residual | 26 | .11 | | | | |
| | Constant | | | | | | 8.674 | |

TABLE B2

STEPWISE REGRESSION FOR TRIAL TWO: ALL VARIABLES

| Step | Variable | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| | | Source | df | MS | $F$ | $R^2$ | $b$ | Beta |
| 1 | SIMHRS | Regression | 1 | 63.36 | 8.68 | .213 | - .459 | - .153 |
| | | Residual | 32 | 7.30 | | | | |
| 2 | TIME 2 | Regression | 2 | 44.74 | 6.68 | .301 | 6.463 | 2.154 |
| | | Residual | 31 | 6.69 | | | | |
| 3 | TIME 6 | Regression | 3 | 34.37 | 5.32 | .347 | 7.221 | 2.407 |
| | | Residual | 30 | 6.46 | | | | |
| 4 | TOTTIM | Regression | 4 | 33.58 | 5.98 | .452 | -4.353 | -2.902 |
| | | Residual | 29 | 5.61 | | | | |
| 5 | TIME 3 | Regression | 5 | 48.16 | 23.98 | .811 | .966 | 1.609 |
| | | Residual | 28 | 2.01 | | | | |
| 6 | TIME 4 | Regression | 6 | 47.09 | 87.80 | .951 | -1.415 | - .472 |
| | | Residual | 27 | .54 | | | | |
| 7 | TIME 7 | Regression | 7 | 40.90 | 99.11 | .964 | .253 | .169 |
| | | Residual | 26 | .41 | | | | |
| 8 | STUDY | Regression | 8 | 36.05 | 104.62 | .971 | - .051 | - .119 |
| | | Residual | 25 | .34 | | | | |
| 9 | UPTHRS | Regression | 9 | 32.06 | 90.84 | .971 | - .029 | - .029 |
| | | Residual | 24 | .35 | | | | |
| | Constant | | | | | | -5.501 | |

## TABLE B3

## STEPWISE REGRESSION FOR TRIAL THREE: ALL VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTT!M | Regression | 1 | 17.22 | 4.80 | .130 | - .587 | - .293 |
| | | Residual | 32 | 3.59 | | | | |
| 2 | SUM 1 | Regression | 2 | 13.29 | 3.91 | .201 | -1.101 | -4.405 |
| | | Residual | 31 | 3.40 | | | | |
| 3 | MAXALT | Regression | 3 | 13.10 | 4.24 | .298 | 1.470 | 2.205 |
| | | Residual | 30 | 3.09 | | | | |
| 4 | SUM 3 | Regression | 4 | 18.90 | 9.72 | .573 | .971 | 4.856 |
| | | Residual | 29 | 1.94 | | | | |
| 5 | TIME 3 | Regression | 5 | 21.75 | 26.22 | .824 | - .429 | -1.502 |
| | | Residual | 28 | .93 | | | | |
| 6 | TOTHRS | Regression | 6 | 20.21 | 50.74 | .919 | .241 | .361 |
| | | Residual | 27 | .40 | | | | |
| 7 | TIME 5 | Regression | 7 | 18.24 | 109.48 | .967 | - .471 | - .235 |
| | | Residual | 26 | .17 | | | | |
| 8 | TIME 6 | Regression | 8 | 16.15 | 144.34 | .979 | - .281 | - .141 |
| | | Residual | 25 | .11 | | | | |
| 9 | UPTHRS | Regression | 9 | 14.51 | 242.10 | .989 | - .096 | - .144 |
| | | Residual | 24 | .06 | | | | |
| | Constant | | | | | | -1.562 | |

TABLE B4

STEPWISE REGRESSION FOR TRIAL FOUR: ALL VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TIME 2 | Regression | 1 | 61.26 | 27.71 | .464 | 3.570 | 1.785 |
| | | Residual | 32 | 2.21 | | | | |
| 2 | STUDY | Regression | 2 | 47.56 | 39.98 | .721 | - .209 | - .730 |
| | | Residual | 31 | 1.19 | | | | |
| 3 | TOTHRS | Regression | 3 | 34.41 | 35.89 | .782 | - .204 | - .306 |
| | | Residual | 30 | .96 | | | | |
| 4 | SUM 1 | Regression | 4 | 28.02 | 40.75 | .849 | - .617 | -1.235 |
| | | Residual | 29 | .69 | | | | |
| 5 | TIME 3 | Regression | 5 | 24.94 | 95.59 | .945 | .272 | .544 |
| | | Residual | 28 | .26 | | | | |
| 6 | TIME 5 | Regression | 6 | 21.82 | 534.55 | .992 | .475 | .237 |
| | | Residual | 27 | .04 | | | | |
| 7 | TIME 4 | Regression | 7 | 18.78 | 945.88 | .996 | - .159 | - .080 |
| | | Residual | 26 | .02 | | | | |
| 8 | MAXALT | Regression | 8 | 16.47 | 1669.77 | .998 | .025 | .086 |
| | | Residual | 25 | .01 | | | | |
| 9 | TIME 6 | Regression | 9 | 14.65 | 2725.01 | .999 | - .081 | - .041 |
| | | Residual | 24 | .01 | | | | |
| | Constant | | | | | | 12.373 | |

TABLE B5

STEPWISE REGRESSION FOR TRIAL FIVE: ALL VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TIME 2 | Regression | 1 | 99.23 | 16.06 | .334 | 5.430 | 1.810 |
| | | Residual | 32 | 6.18 | | | | |
| 2 | TIME 3 | Regression | 2 | 75.64 | 16.09 | .509 | -1.330 | - .443 |
| | | Residual | 31 | 4.70 | | | | |
| 3 | MAXALT | Regression | 3 | 64.80 | 16.94 | .655 | 1.508 | 1.005 |
| | | Residual | 30 | 3.42 | | | | |
| 4 | SUM 2 | Regression | 4 | 54.68 | 20.26 | .736 | - .338 | - .225 |
| | | Residual | 29 | 2.70 | | | | |
| 5 | STUDY | Regression | 5 | 45.91 | 19.06 | .773 | - .253 | - .591 |
| | | Residual | 28 | 2.41 | | | | |
| 6 | SUM 1 | Regression | 6 | 41.48 | 23.28 | .838 | -2.443 | -1.629 |
| | | Residual | 27 | 1.78 | | | | |
| 7 | TOTTIM | Regression | 7 | 36.66 | 23.61 | .864 | 1.625 | .542 |
| | | Residual | 26 | 1.55 | | | | |

Table B5 (continued)

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 8 | TIME 7 | Regression | 8 | 33.71 | 30.85 | .908 | 2.762 | .921 |
| | | Residual | 25 | 1.09 | | | | |
| 9 | UPTHRS | Regression | 9 | 31.56 | 58.63 | .957 | .403 | .403 |
| | | Residual | 24 | .54 | | | | |
| 10 | TIME 4 | Regression | 10 | 28.92 | 85.72 | .974 | .591 | .197 |
| | | Residual | 23 | .34 | | | | |
| 11 | TOTHRS | Regression | 11 | 26.30 | 75.12 | .974 | .017 | .017 |
| | | Residual | 22 | .35 | | | | |
| | Constant | | | | | | 5.367 | |

TABLE B6

STEPWISE REGRESSION FOR TRIAL SIX: ALL VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTTIM | Regression | 1 | 248.12 | 13.76 | .301 | 4.429 | .886 |
| | | Residual | 32 | 18.03 | | | | |
| 2 | TIME 3 | Regression | 2 | 225.47 | 18.69 | .547 | 2.490 | 1.992 |
| | | Residual | 31 | 12.07 | | | | |
| 3 | TIME 5 | Regression | 3 | 237.48 | 63.29 | .864 | -5.235 | -1.047 |
| | | Residual | 30 | 3.75 | | | | |
| 4 | SUM 1 | Regression | 4 | 192.92 | 104.92 | .935 | - .930 | - .744 |
| | | Residual | 29 | 1.84 | | | | |
| 5 | TOTHRS | Regression | 5 | 161.92 | 294.60 | .981 | .312 | .187 |
| | | Residual | 28 | .55 | | | | |
| 6 | STUDY | Regression | 6 | 135.95 | 395.84 | .989 | .093 | .130 |
| | | Residual | 27 | .34 | | | | |
| 7 | SUM 2 | Regression | 7 | 117.83 | 18485.00 | 1.000 | - .398 | - .159 |
| | | Residual | 26 | .01 | | | | |
| | Constant | | | | | | - .012 | |

APPENDIX C

TABLE C1

FIVE STEP REGRESSION FOR TRIAL ONE: SEVEN VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | SUM 3 | Regression | 1 | 89.98 | 13.91 | .303 | .131 | .612 |
| | | Residual | 32 | 6.47 | | | | |
| 2 | SUM 1 | Regression | 2 | 47.00 | 7.18 | .316 | -.125 | -.250 |
| | | Residual | 31 | 6.55 | | | | |
| 3 | TOTTIM | Regression | 3 | 32.73 | 4.94 | .331 | .216 | .144 |
| | | Residual | 30 | 6.63 | | | | |
| 4 | TIME 2 | Regression | 4 | 25.08 | 3.70 | .338 | .222 | .074 |
| | | Residual | 29 | 6.78 | | | | |
| 5 | MAXALT | Regression | 5 | 20.35 | 2.92 | .343 | .030 | .079 |
| | | Residual | 28 | 6.97 | | | | |
| | Constant | | | | | | 4.714 | |

TABLE C2

FIVE STEP REGRESSION FOR TRIAL TWO: SEVEN VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | $F$ | $R^2$ | $b$ | Beta |
| 1 | TIME 2 | Regression | 1 | 52.70 | 6.90 | .177 | 3.746 | 1.249 |
| | | Residual | 32 | 7.63 | | | | |
| 2 | SUM 1 | Regression | 2 | 33.99 | 4.60 | .229 | -1.332 | -2.664 |
| | | Residual | 31 | 7.39 | | | | |
| 3 | TIME 3 | Regression | 3 | 38.17 | 6.27 | .386 | .801 | 1.335 |
| | | Residual | 30 | 6.08 | | | | |
| 4 | SUM 3 | Regression | 4 | 40.37 | 8.64 | .544 | .263 | .876 |
| | | Residual | 29 | 4.67 | | | | |
| 5 | MAXALT | Regression | 5 | 32.70 | 6.86 | .550 | - .054 | - .108 |
| | | Residual | 28 | 4.77 | | | | |
| | Constant | | | | | | 4.117 | |

TABLE C3

FIVE STEP REGRESSION FOR TRIAL THREE: SEVEN VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTTIM | Regression<br>Residual | 1<br>32 | 17.22<br>3.59 | 4.80 | .130 | .156 | .078 |
| 2 | SUM 1 | Regression<br>Residual | 2<br>31 | 13.29<br>3.40 | 3.91 | .201 | - .827 | -3.307 |
| 3 | MAXALT | Regression<br>Residual | 3<br>30 | 13.10<br>3.09 | 4.24 | .298 | - 1.086 | 1.629 |
| 4 | SUM 3 | Regression<br>Residual | 4<br>29 | 18.90<br>1.94 | 9.72 | .573 | .709 | 3.543 |
| 5 | TIME 3 | Regression<br>Residual | 5<br>28 | 21.75<br>.83 | 26.22 | .824 | - .391 | -1.369 |
| | Constant | | | | | | - .823 | |

TABLE C4

FIVE STEP REGRESSION FOR TRIAL FOUR: SEVEN VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TIME 2 | Regression | 1 | 61.26 | 27.71 | .464 | 1.554 | .777 |
| | | Residual | 32 | 2.21 | | | | |
| 2 | MAXALT | Regression | 2 | 35.29 | 17.80 | .535 | .104 | .365 |
| | | Residual | 31 | 1.98 | | | | |
| 3 | SUM 3 | Regression | 3 | 24.40 | 12.45 | .555 | - .060 | - .181 |
| | | Residual | 30 | 1.96 | | | | |
| 4 | TOTTIM | Regression | 4 | 18.89 | 9.71 | .573 | .324 | .162 |
| | | Residual | 29 | 1.95 | | | | |
| 5 | SUM 1 | Regression | 5 | 15.19 | 7.58 | .575 | - .064 | - .128 |
| | | Residual | 28 | 2.00 | | | | |
| | Constant | | | | | | 3.615 | |

TABLE C5

FIVE STEP REGRESSION FOR TRIAL FIVE: SEVEN VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | *F* | $R^2$ | *b* | Beta |
| 1 | TIME 2 | Regression | 1 | 93.23 | 16.06 | .334 | 2.363 | .788 |
| | | Residual | 32 | 6.18 | | | | |
| 2 | TIME 3 | Regression | 2 | 75.64 | 16.09 | .509 | -1.353 | -.451 |
| | | Residual | 31 | 4.70 | | | | |
| 3 | MAXALT | Regression | 3 | 64.80 | 18.94 | .655 | .737 | .492 |
| | | Residual | 30 | 3.42 | | | | |
| 4 | SUM 2 | Regression | 4 | 54.68 | 20.26 | .736 | .679 | .453 |
| | | Residual | 29 | 2.70 | | | | |
| 5 | SUM 1 | Regression | 5 | 45.65 | 18.59 | .768 | - .564 | - .376 |
| | | Residual | 28 | 2.46 | | | | |
| | Constant | | | | | | 2.203 | |

TABLE C6

FIVE STEP REGRESSION FOR TRIAL SIX: SEVEN VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTTIM | Regression | 1 | 248.12 | 13.76 | .301 | 3.570 | .714 |
| | | Residual | 32 | 18.03 | | | | |
| 2 | TIME 3 | Regression | 2 | 225.47 | 18.69 | .547 | 1.214 | .971 |
| | | Residual | 31 | 12.07 | | | | |
| 3 | SUM 3 | Regression | 3 | 158.33 | 13.57 | .576 | - .929 | -.929 |
| | | Residual | 30 | 11.67 | | | | |
| 4 | MAXALT | Regression | 4 | 124.44 | 11.03 | .603 | .367 | .220 |
| | | Residual | 29 | 11.28 | | | | |
| 5 | SUM 2 | Regression | 5 | 103.32 | 9.38 | .626 | .850 | .340 |
| | | Residual | 28 | 11.01 | | | | |
| | Constant | | | | | | 2.226 | |

APPENDIX D

TABLE D1

FIVE STEP REGRESSION FOR TRIAL ONE: NINE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | $F$ | $R^2$ | $b$ | Beta |
| 1 | SUM 3 | Regression | 1 | 89.98 | 13.91 | .303 | 1.092 | 5.095 |
| | | Residual | 32 | 6.47 | | | | |
| 2 | TOTHRS | Regression | 2 | 70.14 | 13.87 | .472 | - .886 | - .886 |
| | | Residual | 31 | 5.06 | | | | |
| 3 | TIME 3 | Regression | 3 | 61.68 | 16.53 | .623 | -1.453 | -2.422 |
| | | Residual | 30 | 3.73 | | | | |
| 4 | SUM 2 | Regression | 4 | 62.07 | 36.94 | .836 | - .866 | -2.887 |
| | | Residual | 29 | 1.68 | | | | |
| 5 | MAXALT | Regression | 5 | 51.44 | 36.17 | .866 | .071 | .190 |
| | | Residual | 28 | 1.42 | | | | |
| | Constant | | | | | | 6.210 | |

TABLE D2

FIVE STEP REGRESSION FOR TRIAL TWO; NINE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TIME 2 | Regression | 1 | 52.70 | 6.90 | .177 | 6.768 | 2.256 |
| | | Residual | 32 | 7.63 | | | | |
| 2 | SUM 1 | Regression | 2 | 33.99 | 4.60 | .229 | -2.514 | -5.029 |
| | | Residual | 31 | 7.39 | | | | |
| 3 | TIME 3 | Regression | 3 | 38.17 | 6.27 | .386 | 1.350 | 2.250 |
| | | Residual | 30 | 6.08 | | | | |
| 4 | SUM 3 | Regression | 4 | 40.37 | 8.64 | .544 | .413 | 1.378 |
| | | Residual | 29 | 4.67 | | | | |
| 5 | STUDY | Regression | 5 | 52.47 | 42.40 | .883 | - .368 | - .859 |
| | | Residual | 28 | 1.24 | | | | |
| | Constant | | | | | | 15.105 | |

TABLE D3

FIVE STEP REGRESSION FOR TRIAL THREE: NINE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | *F* | $R^2$ | *b* | Beta |
| 1 | TOTTIM | Regression | 1 | 17.22 | 4.80 | .130 | .156 | .078 |
| | | Residual | 32 | 3.59 | | | | |
| 2 | SUM 1 | Regression | 2 | 13.29 | 3.91 | .201 | - .827 | -3.307 |
| | | Residual | 31 | 3.40 | | | | |
| 3 | MAXALT | Regression | 3 | 13.10 | 4.23 | .298 | 1.086 | 1.629 |
| | | Residual | 30 | 3.09 | | | | |
| 4 | SUM 3 | Regression | 4 | 18.90 | 9.72 | .573 | .709 | 3.543 |
| | | Residual | 29 | 1.94 | | | | |
| 5 | TIME 3 | Regression | 5 | 21.75 | 26.22 | .824 | - .391 | -1.369 |
| | | Residual | 28 | .83 | | | | |
| | Constant | | | | | | - .823 | |

TABLE D4

FIVE STEP REGRESSION FOR TRIAL FOUR: NINE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | $F$ | $R^2$ | $b$ | Beta |
| 1 | TIME 2 | Regression | 1 | 61.26 | 27.71 | .464 | 3.556 | 1.778 |
| | | Residual | 32 | 2.21 | | | | |
| 2 | STUDY | Regression | 2 | 47.56 | 39.98 | .721 | - .207 | - .725 |
| | | Residual | 31 | 1.19 | | | | |
| 3 | TOTHRS | Regression | 3 | 34.41 | 35.89 | .782 | - .227 | - .340 |
| | | Residual | 30 | .96 | | | | |
| 4 | SUM 1 | Regression | 4 | 28.02 | 40.75 | .849 | - .626 | -1.252 |
| | | Residual | 29 | .69 | | | | |
| 5 | TIME 3 | Regression | 5 | 24.94 | 95.59 | .945 | .352 | .703 |
| | | Residual | 28 | .26 | | | | |
| | Constant | | | | | | 12.572 | |

TABLE D5

FIVE STEP REGRESSION FOR TRIAL FIVE: NINE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | *F* | $R^2$ | *b* | Beta |
| 1 | TIME 2 | Regression | 1 | 99.23 | 16.06 | .334 | 1.593 | .531 |
| | | Residual | 32 | 6.18 | | | | |
| 2 | TIME 3 | Regression | 2 | 75.64 | 16.09 | .509 | -1.763 | -.588 |
| | | Residual | 31 | 4.70 | | | | |
| 3 | MAXALT | Regression | 3 | 64.80 | 18.94 | .655 | .704 | .470 |
| | | Residual | 30 | 3.42 | | | | |
| 4 | SUM 2 | Regression | 4 | 54.68 | 20.26 | .736 | .523 | .349 |
| | | Residual | 29 | 2.70 | | | | |
| 5 | STUDY | Regression | 5 | 45.91 | 19.06 | .773 | - .088 | -.206 |
| | | Residual | 28 | 2.41 | | | | |
| | Constant | | | | | | 4.091 | |

TABLE D6

FIVE STEP REGRESSION FOR TRIAL SIX: NINE VARIABLES

| | | Analysis of Variance | | | | Summary Values | | |
|---|---|---|---|---|---|---|---|---|
| Step | Variable | Source | df | MS | F | $R^2$ | b | Beta |
| 1 | TOTTIM | Regression | 1 | 248.12 | 13.76 | .301 | 3.276 | .655 |
| | | Residual | 32 | 18.03 | | | | |
| 2 | TIME 3 | Regression | 2 | 225.47 | 18.69 | .547 | .728 | .582 |
| | | Residual | 31 | 12.07 | | | | |
| 3 | TOTHRS | Regression | 3 | 161.20 | 14.17 | .586 | .265 | .159 |
| | | Residual | 30 | 11.38 | | | | |
| 4 | MAXALT | Regression | 4 | 126.18 | 11.42 | .612 | .397 | .238 |
| | | Residual | 29 | 11.04 | | | | |
| 5 | SUM 3 | Regression | 5 | 102.81 | 9.26 | .623 | - .265 | -.265 |
| | | Residual | 28 | 11.11 | | | | |
| | Constant | | | | | | -1.870 | |